APPENDIX B. STATISTICAL ANALYSES

This page intentionally left blank.

Table of Contents

Tables

Figures

Acronyms

95URL	95% upper confidence limit (on the mean)
ANOVA	analysis of variance
COC	contaminant of concern
cPAH	carcinogenic polycyclic aromatic hydrocarbon
CV	coefficient of variation
DQO	data quality objective
dw	dry weight
EAA	early action area
ECDF	empirical cumulative distribution function
Ecology	Washington State Department of Ecology
EPA	US Environmental Protection Agency
GOF	goodness-of-fit
HHRA	human health risk assessment
ID	identification
IQR	interquartile range
LDW	Lower Duwamish Waterway
MDD	minimum detectable difference
MDL	method detection limit
OLS	ordinary least squares
OSV	ocean survey vessel
PCB	polychlorinated biphenyl
PE	polyethylene
PPCC	probability plot correlation coefficient
QAPP	quality assurance project plan
RAL	remedial action level
RAO	remedial action objective
RI	remedial investigation
RL	reporting limit

RM	river mile
SCO	sediment cleanup objective
SD	standard deviation
SE	standard error (of the mean)
TEQ	toxic equivalent
TTL	target tissue level
VCA	variance components analysis

This page intentionally left blank.

B1 Introduction and Methods

This appendix to the *Lower Duwamish Waterway Data Evaluation Report* presents the analytical methods and detailed results of the statistical evaluations that were used to interpret the baseline datasets in light of the data quality objectives (DQOs).

The remainder of this appendix is organized into sections that parallel the structure of the main report, which includes the following sections:

- Section 2 – Sediment
- Section 3 – Surface Water
- Section 4 – Fish and Crab Tissue
- Section 5 – Clam Tissue
- Section 6 – References

The statistical methods that were applied to one or more datasets in later sections are described in Section 1.

B1.1 Choosing the Distributional Form for Calculating 95UCLs

The 95% upper confidence limit (on the mean) (95UCL) is a summary statistic required for many of the Lower Duwamish Waterway (LDW) baseline datasets. The 95UCL for each dataset was calculated using the appropriate parametric equations following identification of the most appropriate distributional form (i.e., normal, log-normal, or gamma). If one of the parametric distributions was not appropriate for a dataset, then a non-parametric approach was required. This process also allowed for identification of any possible outliers in a dataset so that these elevated values could be discussed further.

Each dataset was evaluated using tools in ProUCL 5.1 (EPA 2016) and select packages (e.g., EnvStats (Millard 2013) and ggplot2 (Wickham 2009)) in R (R Core Team 2018)). The statistical tools used during this assessment included probability plots, distributional goodness-of-fit (GOF) tests, and graphical and formal outlier tests.

B1.1.1 Goodness-of-fit test

A formal GOF test was conducted for each chemical dataset individually, and each test was confirmed by patterns observed in the probability plots (discussed below). The best-fitting distribution was identified as the one that passed the GOF test and had the highest probability plot correlation coefficient (PPCC). If no distributions provided a reasonable fit to the data, then non-parametric estimates for the 95UCL were required.

For this evaluation, GOF testing relied on the significance of the probability plot correlation coefficient (using *EnvStats::gofTest(x, test="ppcc," estimate.params=TRUE)* in R) for normal, lognormal, and gamma distributions, with the hypothesized

distribution rejected when $p < 0.05$. Once the best distributional fit for a dataset was identified, the 95UCL was calculated in ProUCL 5.1 (EPA 2016).

B1.1.2 Probability plots

Probability plots show the observed quantiles for the dataset on the y-axis vs. the expected quantiles under the theorized distribution on the x-axis (hence the synonymous name "QQ Plot," which stands for quantile-quantile plot). If the theoretical distribution is a reasonable description for the dataset, then this plot should follow an approximately straight line. The best-fit regression line is added to the plots to facilitate interpretation of the GOF indicated by these plots. These plots are generated in R using the function *EnvStats::qqPlot(x, estimate.params=TRUE)*. The presence of potential outliers and systematic deviations from the theorized distribution can also be observed on these plots; if present, such outliers and deviations may lead to a formal outlier test, as described in the next section. Figure B1-1 shows example probability plots for a skewed dataset that is poorly described by a normal distribution (i.e., the observed quantiles do not fit a straight line when plotted against the normal quantiles) but adequately described by a lognormal distribution (i.e., the QQ plot follows an approximately straight line).

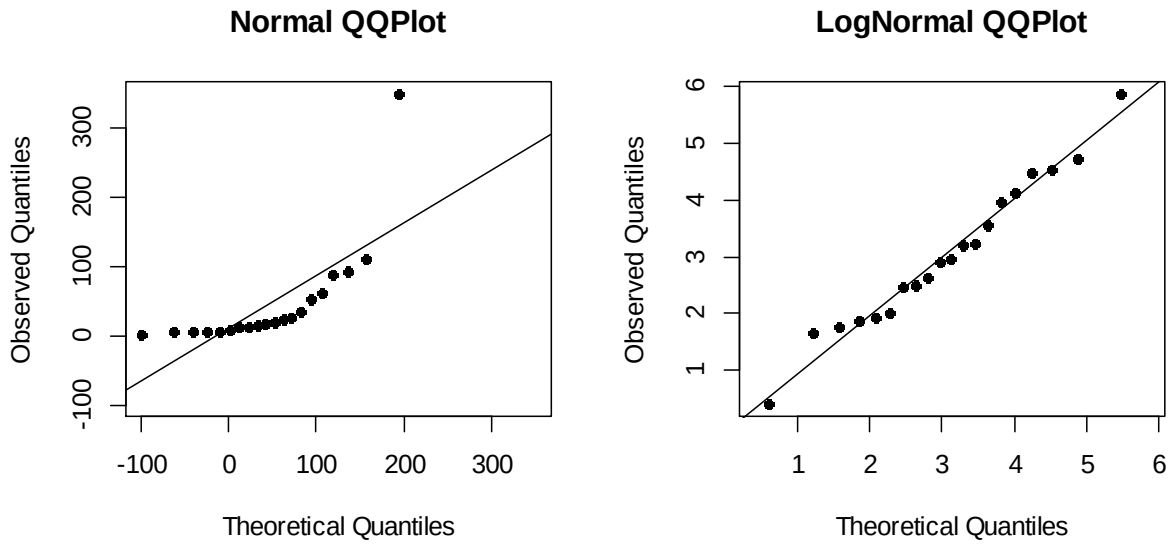


Figure B1-1. Example probability plots for a skewed dataset that does not follow a normal distribution (left) but does follow a lognormal distribution (right)

B1.1.3 Outlier tests

The presence of potential outliers was identified initially through visual inspection of the probability plots. When data points appeared to be extreme (either at the high or low end), a formal outlier test was used. Outlier tests require a parametric assumption for the underlying data; there is no such thing as an outlier for a non-parametric distribution. The two outlier tests used are based on an underlying normal distribution. It is usually the case that the skewness introduced by extreme values can be adequately described by a log-normal or gamma distribution. Alternatively, once extreme values have been removed, the data may be adequately described by a normal distribution, which is the basis for the two outlier tests: Dixon's ($n < 25$, single outliers only) and Rosner's ($n \geq 25$, multiple outliers). Both tests were applied using tools in ProUCL 5.1 (EPA 2016).

B1.1.4 Boxplots

Boxplots (a.k.a. box-and-whisker plots) are used to illustrate the distribution of the data, providing information about the location and spread of the data as well as skewness. Boxplots are especially useful when several are placed side by side. Each boxplot has a shaded/colored rectangle that shows the spread of values between the 1st and 3rd quartiles (i.e., the 25th and 75th percentiles). The height of this rectangle is the interquartile range (IQR), which is simply the value of the third quartile minus the value of the first quartile. The line inside the box indicates the median and the blue diamond indicates the mean; the outer brackets (the "whiskers") represent the minimum and maximum values or 1.5 times the IQR from the median, whichever is less; values outside the whiskers are possible extreme values and are shown as individual data points. The median plus and minus 1.5 times the IQR is expected to contain about 98% of a Standard Normal (Gaussian) distribution. Boxplots were generated in R using the function *ggplot* + *geom_boxplot*. Figure B1-2 is an example boxplot with labels of the distributional characteristics represented by the different parts of the boxplot.

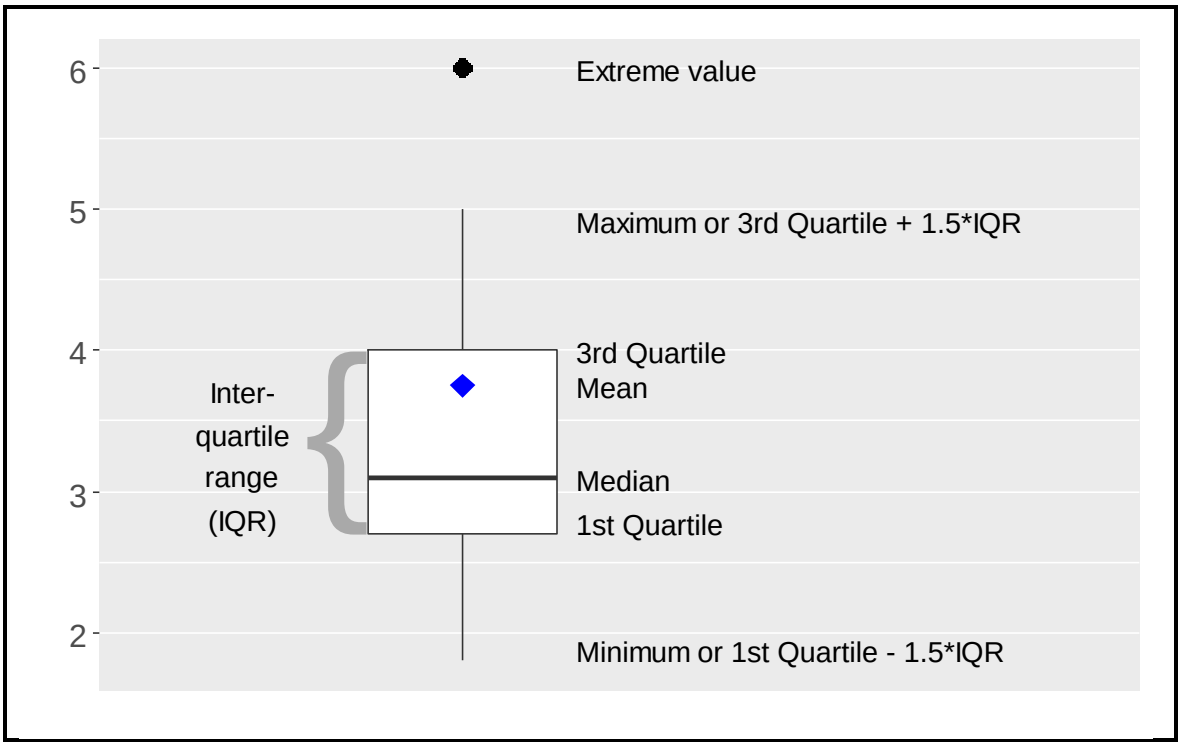
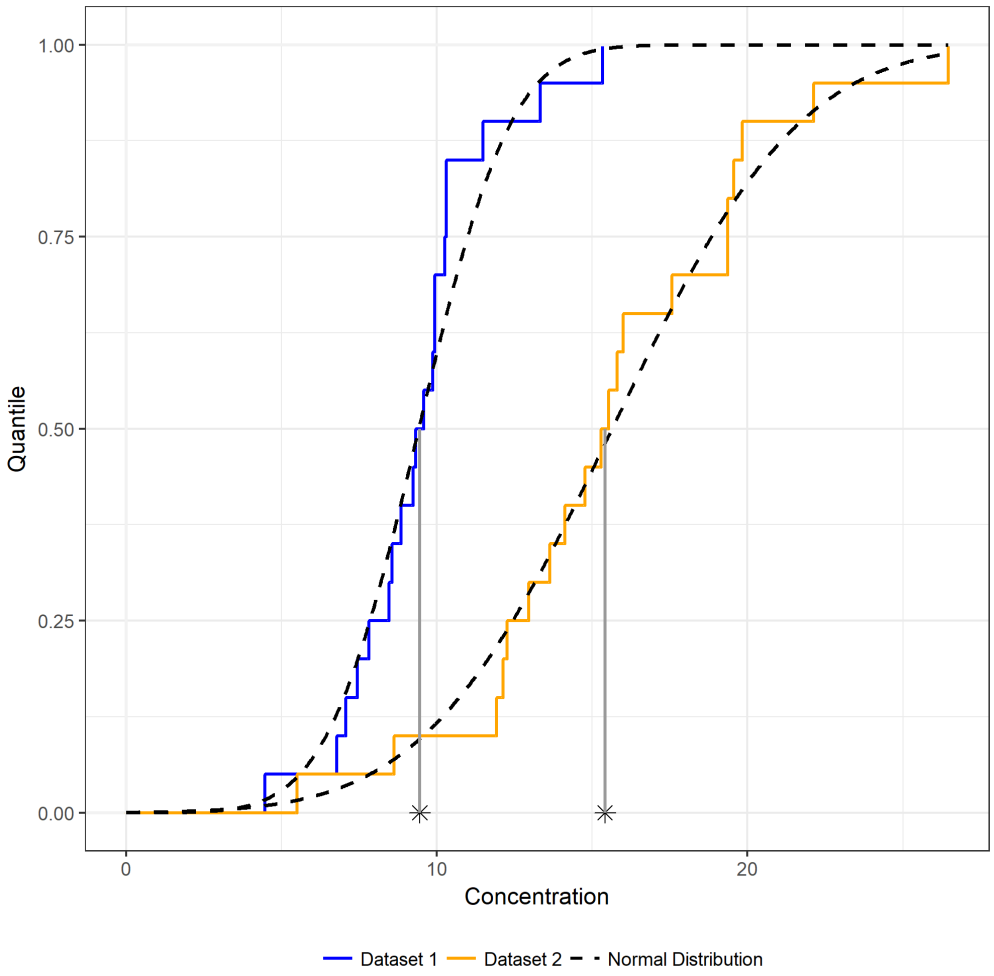


Figure B1-2. Example boxplot with labels of the distributional characteristics represented by the different parts of the boxplot

B1.2 EMPIRICAL CUMULATIVE DISTRIBUTION PLOTS

Useful in illustrating the distribution and skewness in a dataset, empirical cumulative distribution function (ECDF) plots display the percentiles or cumulative probabilities for each observation in a dataset (Figure B1-3). Each distribution is shown as a step function, with a step up at each unique concentration. These plots provide visualization of where an individual threshold concentration (e.g., a cleanup level) may fall along the distribution of concentrations in the LDW dataset, and can also be used to compare multiple datasets (e.g., in cases where the remedial action objective [RAO] cleanup levels are background based, the ECDF for the LDW baseline dataset can be contrasted with the ECDF for the ocean survey vessel (OSV) *Bold* background dataset (Figure B2-1) (DMMP 2009). ECDF plots readily allow the interpretation of certain characteristics of data distributions. When two curves are shown on the same plot, the curve further to the right has higher concentrations (e.g., Dataset 2 in Figure B1-3); steeper curves have less variance (e.g., Dataset 1 in Figure B1-3) and specific percentiles can be readily identified (e.g., the median concentration is the concentration on the x-axis that coincides with a cumulative probability of 0.5 on the y-axis for a particular curve, as indicated by asterisks on the x-axis in Figure B1-3). ECDF plots were generated in R using the function *ggplot + stat_ecdf.*



Note: Asterisks indicate the median concentration for the two datasets. The normal distribution curves use mean and variance estimated from each dataset.

Figure B1-3. Example ECDF plot

B1.3 VARIANCE COMPONENTS ANALYSIS

An analysis of the variance components was used to investigate the relative importance of different sources of variability within the total sampling variance (i.e., small- or large-scale spatial variability). This analysis was applied to the intertidal sediment datasets from Beach Play Areas 1 and 6 and the passive sampler polychlorinated biphenyl (PCB) dataset for surface water.

The intertidal sediment beach play area datasets had two field replicates nested within each of the three composites representing Beach Play Areas 1 and 6; however, the low level of replication introduces uncertainty regarding the general applicability of these results.

The passive sampler PCB dataset was a crossed design (i.e., two stations in each of the two years) with nine replicate samplers in each location-year combination, except for station PS1 in 2018, which had one replicate rejected. This slight imbalance was

corrected by using the average of the other eight replicates in place of the rejected replicate at station PS1 in 2018. The variance components analysis (VCA) was conducted on the balanced dataset.

Even though these datasets were not explicitly designed for this analysis, the results are suggestive of possible patterns in the underlying data populations. These patterns identify possible sources of variance that may be explicitly tested in future monitoring events, during which efforts can be made to reduce that variance if it is meaningful.

VCA uses the analysis of variance (ANOVA) model to partition the sums of squares into their component parts. In the same way that the sources of variance in an ANOVA model are isolated in hypothesis testing to express the "statistical significance" of a particular feature of the study design, the sources of variance can be expressed as a percent of the total to express the relative importance of the variance of that feature. The VCAs were conducted using *anovaVCA* in R (Schuetzenmeister and Dufey 2018). A conservative estimate of total variance was used by setting negative variance components estimates to zero *(NegVC = FALSE)*. Negative variance components can arise from the additive model used to partition the sums of squares from the expected variance components, described below.

Table B1-1 shows the theoretical (modeled) expectation for each variance component in a balanced design with multiple levels of one factor that represents a source of variance (e.g., location, or year) and replication within each level of that factor (e.g., field replicates of the beach play composites, or polyethylene (PE) sampler replication at each station within each year). The expected variance components are derived from the observed mean squares by subtraction.

Table B1-1. Expected means squares for VCA of a single (random) factor and one level of replication in a balanced design

Source	Degrees of Freedom	Expected Mean Squares	Observed Mean Squares	Estimated Variance Component
Factor A (e.g., location)	*a* – 1	$\sigma_\varepsilon^2 + n\sigma_A^2$	SSA/*(a-1)*	$\widehat{\sigma_A^2} = \frac{1}{n}\left[\frac{SSA}{a-1} - \frac{SSE}{a(n-1)}\right]$
Within Factor A (e.g., field replication)	*a*(*n*-1)	σ_ε^2	*SSE/a(n-1)*	$\widehat{\sigma_\varepsilon^2} = \left[\frac{SSE}{a(n-1)}\right]$

SSA – sum of squared residuals for Factor A =

$$SSA = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})^2$$

SSE – sum of squared residual errors =

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

Where *i* indicates the level of Factor A (running from 1 to a), and *j* indicates the individual observation within each level of Factor A (running from 1 to n).

VCA – variance components analysis

For example, in a design that has only one factor with replication, the expected mean squares among the lowest level of replication (e.g., field replicates for the beach play composites) are the estimates of error variance (σ_ε^2). The expected mean squares among the primary factor levels of Factor A (e.g., the independent beach play composites) are the sum of the independent sources of: a) error variance (σ_ε^2) and b) variance among levels of the primary factor ($n\sigma_A^2$) (where n is the number of replicates within each level of the factor, and σ_A^2 is the variance among those levels). The mean squared error among levels of Factor A provides an estimate of the total from that variance source (i.e., $\sigma_\varepsilon^2 + n\sigma_S^2$). Hence, the variance due to sampling location only (σ_S^2) is estimated by subtracting σ_ε^2 and dividing by n. This can lead to negative estimates for a variance component (a mathematical possibility but an odd situation, nonetheless) if there is high variability in σ_ε^2, and if the values from the higher level of replication (sampling locations) all overlap with one another.

B2 Sediment

This section provides details of the statistical analyses summarized in Section 2 of the main report for the sediment data collected in February/March and June 2018 per the surface sediment quality assurance project plan (QAPP) (Windward 2018a). The data were presented in the sediment data report (Windward 2018b).

B2.1 Composite Surface Sediment (0–10-cm) Samples

B2.1.1 Data summary

The surface sediment composite sample dataset consisted of 24 samples; each composite sample was composed of 7 grab samples. Composite samples were analyzed for the four risk drivers (total PCB Aroclors, carcinogenic polycyclic aromatic hydrocarbons [cPAHs], dioxins/furans, and arsenic).

All four risk drivers had 100% detection frequency.[1] Total PCBs were calculated as the sum of detected Aroclors; at least one Aroclor was detected in each sample. Every cPAH compound in each composite sample was detected. For dioxins/furans, non-detected congeners, when included at ½ detection limit, represented ≤ 1% of the total TEQ for most of the samples; in the remaining 6 samples, non-detected congener TEQ contributions ranged from 2 to 34%. The minor presence of non-detects in this dataset did not negatively affect the utility of these data to estimate site-wide mean and 95UCL estimates.

[1] Total PCBs (sum of Aroclors) was a sum of detected values only. If no Aroclors were detected, then the sum was reported as the highest reporting limit (RL) and U-qualified (not detected at given concentration). For weighted sums (i.e., toxic equivalents [TEQs]), non-detects were included at one-half the RL. If none of the components were detected, the sum of the weighted one-half RLs was reported (and TEQ was U-qualified).

B2.1.2 95UCL calculations

Sediment DQO 1 required that the 95UCL for the site-wide mean be established from this dataset for the four risk drivers. Following the methods described in Section 1.1, the best distributional form for each contaminant of concern (COC) was identified and the 95UCL was calculated in ProUCL 5.1 (Table 2-2 in main report). GOF and variance summary statistics for the four risk drivers in the composite sediment dataset are shown in Table B2-1 and illustrated in the probability plots (Figures B2-1a, B2-1b).

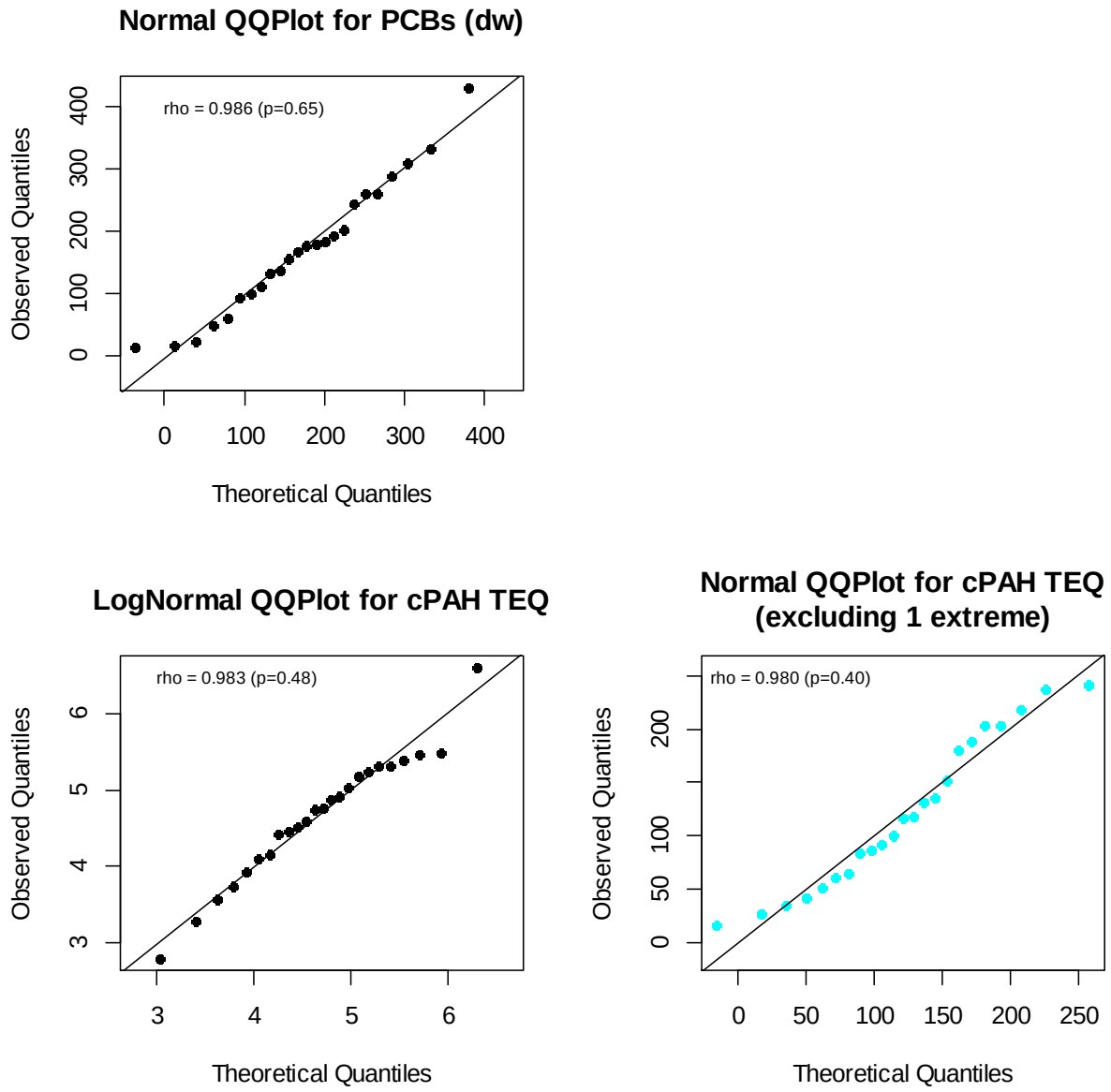
Table B2-1. Goodness-of-fit and variance summary statistics for COCs in 0–10-cm sediment composite samples

COC (units)	Best-fitting Distribution	PPCC[a] (p-value)	CV	Comment
PCB sum of Aroclors (ug/kg dw)	normal	0.986 (p = 0.65)	0.62	The normal distribution is a good fit.
cPAH TEQ (µg/kg dw)	lognormal	0.983 (p = 0.48)	0.98	One elevated influential value present, which skews the dataset.
cPAH TEQ (µg/kg) – exclude outlier	normal	0.98 (p = 0.40)	0.58	Distribution excludes highest value (COMP-02, with concentration of 742 µg/kg). The normal distribution is a good fit.
Dioxin/furan TEQ (ng/kg dw)	gamma	0.986 (p = 0.62)	0.79	Two elevated influential values present, which skew the dataset.
Dioxin/furan TEQ (ng/kg dw) - exclude outliers	normal	0.970 (p=0.18)	0.62	Distribution excludes highest values (COMP-6 and COMP-11, with concentrations of 22.5 and 27.7 ng/kg, respectively). The normal distribution is a good fit.
Arsenic (mg/kg, dw)	lognormal	0.978 (p=0.32)	0.37	One elevated influential value present, which skews the dataset.
Arsenic (mg/kg dw) - exclude outlier	normal	0.994 (p=0.98)	0.26	Distribution excludes highest value (COMP-20, with concentration of 27.2 mg/kg). The normal distribution is a good fit.

[a] PPCC for the best fit distribution for this dataset.

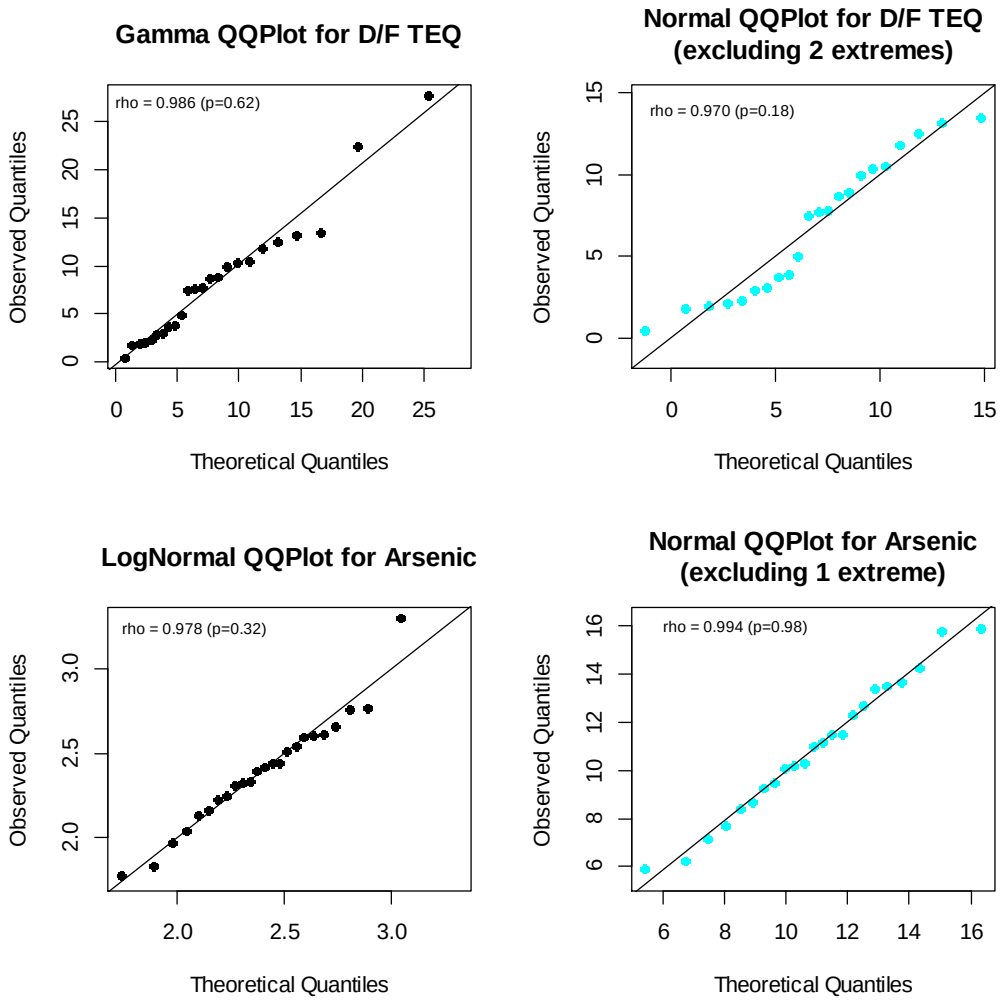
COC – contaminant of concern
CV – coefficient of variation
dw – dry weight
PCB – polychlorinated biphenyl
PPCC – probability plot correlation coefficient
TEQ – toxic equivalent

The data distributions for the cPAH TEQ, dioxin/furan TEQ, and arsenic had one or two elevated influential values each that skewed the data distributions. In order to evaluate the impact of these high values, the best-fit distributions were also evaluated excluding the influential values. When the elevated influential values were excluded (or not present, as in the PCBs dataset), the data distributions were all normally distributed, with coefficients of variation (CVs) of 0.62 or less.



Plots with black dots show the best-fitting distribution for all 24 composite samples. Plots with teal dots show the best-fitting distribution excluding the extreme and influential values that skewed the complete distribution.

Figure B2-1a. Probability plots of total PCB (sum of Aroclors) and cPAH TEQ results in composite samples from 0–10-cm sediments



Plots with black dots show the best-fitting distribution for all 24 composite samples. Plots with teal dots show the best-fitting distribution excluding the extreme and influential values that skewed the distribution of all data.

Figure B2-1b. Probability plots of dioxin/furan TEQ and arsenic results in composite samples from 0–10-cm sediments

With the exception of one or two elevated values, the baseline surface sediment data distributions were all normally distributed, with CVs of approximately 0.6 or less. The composites with elevated values were all collected from areas expected to receive active remediation,[2] hence the distributions excluding these values are expected to be reasonable representations of the sampling variability in site-wide sediments after remediation. For normally distributed datasets, the normal *t*-interval is used to

[2] The composite samples with elevated cPAH and dioxin/furan TEQs are discussed in Section 2.1.1 of the main report. The composite sample with an elevated arsenic concentration, COMP-20, was composed of samples collected between river mile (RM) 3.7 and RM 4.0. This area had five surface sediment samples in the remedial investigation/feasibility study (RI/FS) and post-FS datasets with concentrations that exceeded the arsenic remedial action level (RAL); exceedance factors ranged from 1.4 to 19.

calculate the 95UCL for the site-wide mean. The resulting RME expressed as percent of the mean is calculated using Equation 1:

$$\%\boldsymbol{RME} = \boldsymbol{CV} \times \frac{\boldsymbol{t_{(0.05,n-1)}}}{\sqrt{\mathbf{n}}} \times \mathbf{100} \qquad \textbf{Equation 1}$$

When the CV = 0.6, n = 24 composites, and $\boldsymbol{t_{(0.05,n-1)}}$ = 1.714, the %RME is estimated to be 21% of the mean.

B2.1.3 Arsenic statistical analysis

The arsenic results for the composite sediment dataset were evaluated using an expanded statistical approach because, instead of the arsenic cleanup level being risk-based, it was based on the natural background distribution. The ROD established the arsenic RAO 2 cleanup level as the 95UCL of the dataset collected by the US Environmental Protection Agency's (EPA's) OSV *Bold*, which consisted of 70 individual grab samples collected from Puget Sound natural background areas (EPA 2014). Per the ROD (Table 19, footnote e),[3] determination of compliance may be established by one of the following approaches:

- Approach 1 – A direct comparison of the 95UCL of the LDW dataset mean with the 95UCL of the OSV *Bold* dataset mean (i.e., the background-based cleanup level per the ROD)
- Approach 2 – A statistical comparison of the distribution of the LDW dataset to the OSV *Bold* background dataset

These determinations of compliance can be interpreted as either intending that the post-remedy site should have mean concentrations similar to natural background (Approach 1),[4] or that the entire distribution should be similar to natural background (Approach 2).

There are major differences in the two datasets (i.e., composites from the LDW dataset vs. individual grab samples from Puget Sound in the OSV *Bold* dataset), which influence how compliance, or progress toward compliance, may be appropriately evaluated. For example, the baseline 95UCL based on composite samples may be used to establish whether the site-wide mean can be expected to be below some bright-line threshold with 95% confidence (similar to Approach 1), but the two distributions should not be expected to be similar (Approach 2), because they are of different types of samples (i.e., individual grabs vs. composites). Summary statistics from the two

[3] ROD Table 19 is titled *Cleanup levels for PCBs, arsenic, cPAHs, and dioxins/furans in sediment for human health and ecological COCs (RAOs 1, 2, and 4)*.

[4] Comparing the 95UCL of one distribution to the 95UCL of another distribution does not allow any probability statements to be made about the relationship between the two means. Instead, when this compliance test is met, there will be at least 95% confidence that the post-remedy site mean is less than the bright-line threshold established by the 95UCL of the OSV *Bold* dataset.

datasets (i.e., LDW baseline and OSV *Bold*) were calculated (Table B2-2), and the empirical cumulative distributions of the two datasets are shown in Figure B2-2.

Table B2-2. Summary statistics and test results for arsenic concentrations (mg/kg dw) in the LDW baseline and OSV *Bold* datasets

Dataset	Sample Size	Summary Statistics (mg/kg dw)						PPCC GOF Test	
		Min.	25th Percentile	Median	Mean	75th Percentile	Max.	Corr. Coeff.	p-value
LDW baseline	24	5.90	9.13	11.1	11.6	13.4	27.2	0.978	0.32
OSV *Bold*	70	1.10	3.63	5.95	6.51	8.6	21.0	0.995	0.81
Difference	-	4.8	5.5	5.2	5.1	4.8	6.2	-	-

dw – dry weight
GOF – goodness of fit
LDW – Lower Duwamish Waterway
OSV – ocean survey vessel
PPCC – probability plot correlation coefficient

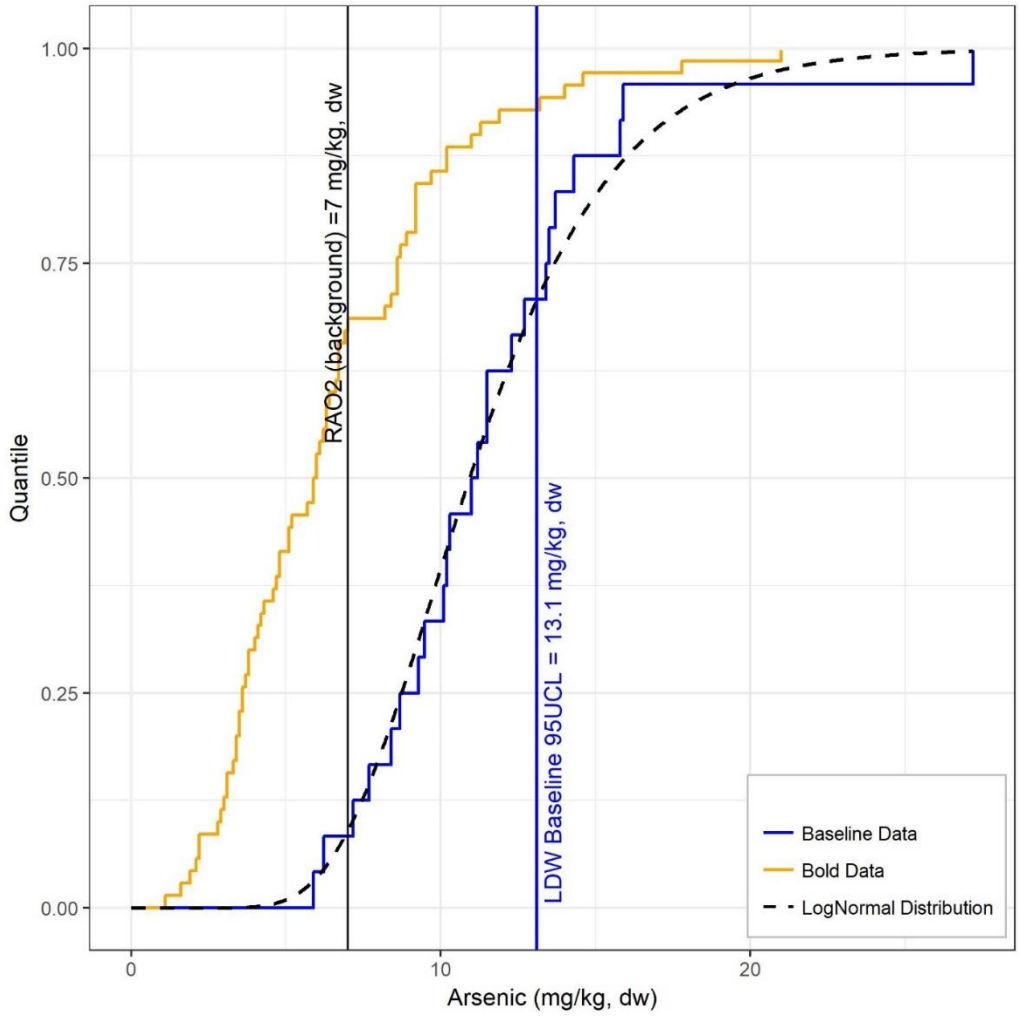


Figure B2-2. Empirical cumulative distribution curves for arsenic concentrations in OSV *Bold* natural background and LDW baseline composite sediment datasets, and theoretical curve for best-fit distribution to the baseline data

The entire distribution of LDW arsenic concentrations is approximately 5 to 6 mg/kg higher than that of the *Bold* dataset. The standard deviations (SDs) were similar at 4.3 mg/kg dry weight (dw) for the LDW dataset and 3.8 mg/kg dw for the OSV *Bold* dataset. Neither of the datasets were significantly different from a lognormal distribution (PPCC GOF test p-values of 0.32 and 0.81 for LDW and *Bold* datasets, respectively; see Section B1.1). Overall, the two distributions showed similar characteristics, but with the LDW distribution shifted to higher concentrations. A distributional comparison between these two datasets indicated that they are statistically different from one another (Kolmogorov-Smirnoff test, $p < 0.001$).

The two approaches identified in the ROD (EPA 2014) are intended to illustrate statistical similarity between the LDW and natural background concentrations, and

both approaches are more restrictive than Washington State cleanup standards [SCUM II] (Ecology 2015).

State standards use the OSV *Bold* natural background dataset differently than does the ROD to establish the sediment cleanup objective (SCO). The Washington State Department of Ecology (Ecology) sets the SCO at a value from the upper tail of the background distribution (the 90/90 upper tolerance limit [UTL][5]), whereas the ROD sets the cleanup level at a value near the central tendency of the background distribution (the 95UCL) (EPA 2014). The subsequent test for compliance within the state standards is to compare the site mean (without a confidence limit) to the background-based 90/90 UTL as a bright-line threshold. This test can be interpreted as requiring concentrations in the post-remedy site population to be, on average and with 50% confidence,[6] not more than concentrations in at least 90% of the natural background population. The baseline arsenic mean is 12 mg/kg, which is in between the background 90/90 UTL of 13 mg/kg based on the OSV *Bold* dataset and the 90/90 UTL of 11 mg/kg based on the OSV *Bold* Plus dataset (Table 10-1, Ecology 2015).

B2.1.4 Power and sample size

Changes in concentrations over time will be evaluated to assess progress toward meeting cleanup goals (DQO 2). In the longer term, if a cleanup level is not met for a given risk driver, its trend may be estimated using a regression analysis (*Pre-Design Studies Work Plan* (Windward and Integral 2017), hereinafter referred to as the Work Plan, Appendix A, Section 2), based on long-term monitoring results. In the near term, this assessment would involve a two-sample, one-tailed comparison between the baseline dataset and a dataset collected post-remediation.

Using the CV results from the Pre-Design Studies, the minimum detectable difference (MDD) between baseline and some future dataset was estimated. The MDD calculations used a Welch's t-test[7] with 24 samples in each dataset, $\alpha = 0.05$, and $\beta \leq 0.10$. Simulations were used to confirm that the MDDs were achieved with at least 90% power.

For each COC that had one or two influential values, the MDD was calculated using two scenarios for future datasets: 1) matching the skewness in the baseline dataset, and 2) matching the baseline dataset without the influential values (i.e., future dataset was normally distributed and had a CV less than baseline). The second scenario is reasonable to use to estimate post-remedy sampling results, since sediment with concentrations greater than the RAL (that were found to be influential in the baseline

[5] The Washington State SCO is the 90/90 UTL of natural background, the 90/90 UTL being the concentration at which there exists 90% confidence that 90% of the natural background population will not exceed the limit.

[6] For a normal distribution, the (1-α)% UCL for the mean is $\overline{\boldsymbol{X}} + \boldsymbol{t}_{(\alpha,\mathbf{df})} \times \boldsymbol{SE}$. A 50% UCL has $\alpha = 0.50$ and a t-value of 0. So the 50% UCL for the mean is equal to the mean.

[7] A two-sample, one-tailed comparison with unequal variances.

dataset) will be remediated. When both baseline and future studies are lognormally distributed, a Welch's t-test on logged data is assumed to be necessary to accommodate the skewness in both datasets. When future studies are normally distributed, the influential values from the baseline dataset do not violate the assumption of normality using simulations, and a Welch's t-test on untransformed data is appropriate.

The MDDs—expressed both as a percent of the baseline mean and as the concentration difference from the baseline mean—were calculated for each COC using the two possible scenarios (Table B2-3). The highest concentration for the mean of a future dataset that would be significantly different from the concentration for the baseline mean is also shown in Table B2-3. These means would have a maximum concentration equal to the baseline mean minus the estimated MDD. For example, PCBs have a baseline mean of 172 µg/kg and an MDD of 62 µg/kg, so a normally distributed dataset with a CV of 0.62 and a mean of (172 – 62) = 110 µg/kg or less would be statistically below baseline ($\alpha = 0.05$, $\beta \leq 0.10$). The baseline design may be expected to detect decreases ranging from 22% of the baseline mean for arsenic to 51% of the baseline mean for cPAH TEQ. The MDD values are slightly smaller using the assumption that the future distribution will have less heterogeneity (matching baseline without influential values).

Table B2-3. Power calculations for comparisons between baseline and future site-wide means in 0–10-cm surface sediments

Chemical	Work Plan CV	Baseline SWAC[a]	Baseline CV (Distr.)	Future Studies CV (Distr.)[b]	Power Calculations (α = 0.05, power = 0.90)		
					MDD[c] as Conc.	MDD as % of Baseline SWAC	Future SWAC Expected to be Significantly Less than Baseline
Total PCBs (Aroclors) (µg/kg dw)	0.7	172	0.62 (N)	0.62 (N)	62 µg/kg	36%	< 110 µg/kg
cPAH TEQ (µg/kg dw)	0.7	147	0.98 (L)	0.98 (L)	75 µg/kg	51%	< 72 µg/kg
				0.58 (N)	68 µg/kg	46%	< 79 µg/kg
Dioxin/furan TEQ (ng/kg dw)	0.7	8.33	0.79 (L)	0.79 (L)	3.7 ng/kg	45%	< 4.6 ng/kg
				0.62 (N)	3.3 ng/kg	40%	< 5.0 ng/kg
Arsenic (mg/kg dw)	0.7	11.6	0.37 (L)	0.37 (L)	2.5 mg/kg	22%	< 9.1 mg/kg
				0.26 (N)	2.4 mg/kg	21%	< 9.2 mg/kg

[a] The arithmetic mean of samples from the baseline survey design is an estimate of the SWAC.

[b] Two different future scenarios are considered: 1) matching the skewness observed in the full baseline dataset, and 2) matching the baseline dataset after the influential data points identified in Table B2-1 have been excluded. N denotes a normal distribution, and L denotes a lognormal distribution.

[c] The MDD calculations used a Welch's two-sample, one-tailed t-test, with $\alpha = 0.05$ and power ≥ 0.9.

CV – coefficient of variation
cPAH – carcinogenic polycyclic aromatic hydrocarbon
dw – dry weight
MDD – minimum detectable difference
PCB – polychlorinated biphenyl
SWAC – spatially weighted average concentration
TEQ – toxic equivalent

The site-wide mean concentrations in the 0–10-cm sediments had 95UCLs that were less than the RAO 2 (netfishing only) cleanup levels for PCBs, cPAH TEQ, and dioxin/furan TEQ (Table 2-2 in main report). Statistical power was calculated using a normal (for PCBs) or lognormal (for cPAH and dioxin/furan TEQs) one-sample, one-tailed t-test to compare the baseline mean to the RAO 2 cleanup level. The statistical power of these comparisons was > 99%.

B2.2 COMPOSITE INTERTIDAL SURFACE SEDIMENT (0–45-CM) SAMPLES

B2.2.1 Potential clamming areas

The dataset for sediments from potential clamming areas consisted of 3 site-wide composite samples with 68 samples in each (for a total of 204 grab samples). Composites were analyzed for PCBs, arsenic, cPAHs, and dioxins/furans.

All four risk drivers had 100% detection frequency in this dataset. The constituent compounds for dioxin/furan TEQ and cPAH TEQ were also 100% detected. Total PCBs were calculated as the sum of detected Aroclors; at least one Aroclor was detected in each sample.

B2.2.1.1 95UCL calculations

Sediment DQO 7 (applies to potential clamming area sediment) requires that the 95UCL for the mean of the LDW-wide potential clamming areas be established from this dataset for the four risk drivers. The 95UCL was derived using a *t*-interval for a normally distributed population[8] using Equation 2.

$$95UCL = \bar{X} + t_{(0.05,df=2)} \times SE \qquad \textbf{Equation 2}$$

Where sample size (n) = 3 and the additional terms are defined as:

$\bar{X}$ = arithmetic mean of the n site-wide composites

SE = standard error calculated as the SD of the n site-wide composites divided by $\sqrt{n}$.

df = degrees of freedom, equal to n - 1

The summary statistics for the four risk drivers in this dataset are presented in Table B2-4 and Figure B2-3. For a small sample size (e.g., n = 3), it is not unexpected that a random sample would appear to be asymmetrical (e.g., total PCBs and dioxins/furans TEQ in the clamming area sediments, Figure B2-3). This apparent skewness does not automatically refute the normality of a small sample size, and the theoretical underpinning of the CLT is relied upon. The CLT states that the mean (i.e., the physical averaging through compositing) of 68 individual samples should be

[8] Because each analytical sample represented the potential clamming area-wide mean based on a large number of grab samples (n = 68) per composite, the Central Limit Theorem (CLT) was invoked and normality assumed.

approximately normally distributed. If the underlying distribution of the composite sample dataset is a skewed distribution rather than a normal distribution (e.g., gamma or lognormal), then the 95UCL provided by Equation 2 will have coverage of the true mean that is less than 95% (i.e., the 95UCL will be too low).

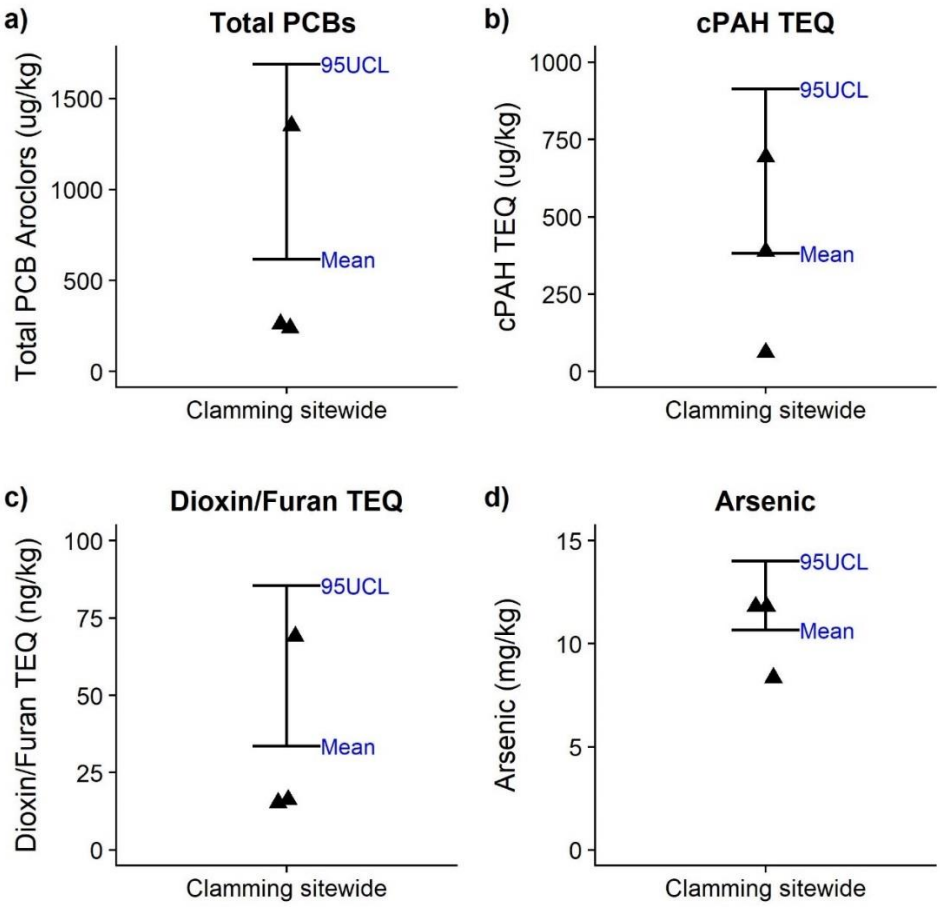
Table B2-4. Summary statistics in potential clamming areas for intertidal (0–45-cm) sediment composites

Sample ID, Summary Statistics	Total PCBs (µg/kg)	cPAH TEQ (µg/kg)	Dioxin/Furan TEQ (ng/kg)	Arsenic (mg/kg)
Composite sample concentrations:				
LDW18-IT45-CL-Comp1	239	388 J	15.3 J	11.8 J
LDW18-IT45-CL-Comp2	1,350 JN	693	69.1 J	11.8 J
LDW18-IT45-CL-Comp3	261 J	61.4	16.3 J	8.35 J
Summary statistics:				
Mean	617	381	33.6	10.7
SD	636	316	30.8	1.99
CV [a]	103%	83%	92%	19%
95UCL[b]	1,690	913	85.5	14.0

[a] CV as % = SD/mean x 100.

[b] 95UCL calculated using the t-interval (degrees of freedom = 2) for the clamming area composites. These estimates do not use the homogenization replicates taken for clamming area composite sample 1 (LDW18-IT45-CL-Comp1). If the homogenization replicates for PCBs and cPAH TEQs were averaged for composite 1 prior to calculating the 95UCL, results would be 1,690 and 878 µg/kg, respectively.

95UCL – 95% upper confidence limit (on the mean)
cPAH – carcinogenic polycyclic aromatic hydrocarbon
CV – coefficient of variation
ID – identification
J – estimated concentration
N – tentative identification
PCB – polychlorinated biphenyl
SD – standard deviation
TEQ – toxic equivalent



Note: For each COC, the three individual sample results are shown (black triangles), along with the mean and 95UCL.

Figure B2-3. Results for the LDW-wide clamming area intertidal (0–45-cm) sediments

B2.2.2 Beach play areas

The 0–45-cm sediment dataset for beach play areas consisted of 24 composite samples (3 composites from each of the 8 beaches, the number of grab samples included in each composite ranging from 3 to 9, based on beach size) plus 6 field replicates. All composite samples were analyzed for PCBs, arsenic, cPAHs, and dioxins/furans.

All four risk drivers had 100% detection frequency, although the constituent compounds for dioxins/furans were not always detected. Total PCBs were calculated as the sum of detected Aroclors;[9] at least one Aroclor was detected in each sample. The individual cPAH compounds were 100% detected in this dataset. For dioxins/furans, non-detected congeners contributed ≤ 1% of the dioxin/furan TEQ for most of the samples; in the remaining 6 samples, the TEQ contributions from non-detects ranged

[9] This is the same summing rule used in the RI/FS.

from 2 to 36%. The samples in which non-detects had the highest percent contribution to the total TEQ were those with the lower total TEQ values**.** The minor presence of non-detects in this dataset did not negatively affect the utility of these data to estimate site-wide mean and 95UCL estimates.

B2.2.2.1 95UCL calculations

Sediment DQO 9 (as applied to beach play area sediments) required that the 95UCL for the mean of each beach be established for the four risk drivers. The 95UCL for the baseline composite samples from each beach was calculated using Chebyshev's inequality (Equation 3). [10]

[10] The shape of the distribution could not be adequately evaluated with only three samples, so a non-parametric Chebyshev interval was used.

$$95UCL = \bar{X} + \sqrt{\left({}^{1}/_{0.05} - 1\right)} \times SE \qquad \textbf{Equation 3}$$

Where the additional term is defined as:

$\bar{X}$ = arithmetic mean of the 3 beach-wide composites

SE = standard error calculated as the SD of the 3 beach-wide composites at each beach, divided by $\sqrt{3}$.

The summary statistics for the four risk drivers in this dataset are presented in Table B2-6. The field replicate samples were used to assess a combination of spatial variance within the sampling locations, homogenization variance, and analytical variance; however, they were not included in the calculation of the beach-wide means or standard errors (SEs) [11] for Beaches 1 and 6. This allowed for similar interpretations of the 95UCL estimates from every beach (i.e., each UCL represented the confidence limit for the mean of three composites).

[11] In the Pre-Design Studies database, primary and field replicate results were retained as discrete samples (Windward and Integral 2017). The field replicates were intended as quality assurance/quality control samples, to be used to evaluate the efficiency of field contamination procedures and the variability attributable to sample handling (sediment QAPP)—hence the decision to use only the primary sample results for baseline summaries (Windward 2018a).

Table B2-6. Overview of composite data and summary statistics in beach play areas for intertidal (0–45-cm) sediments

Beach Play Area	No. of Grab Samples per Composite	Composite Concentrations			Summary Statistics		
		Composite 1	Composite 2	Composite 3	Mean	95UCL[a]	CV
Total PCBs (µg/kg)							
Beach 1	3	265	78.7 J	17.0	120	445	108%
Beach 2	3	120.3 J	118.1	66.2	102	179	30%
Beach 3	5	69.7 J	238.6	23.0 JN	110	396	103%
Beach 4	5	322 J	556 JN	199.5	359	815	51%
Beach 5	9	92.4 JN	160.2 J	90.4 JN	114	214	35%
Beach 6	3	184	990 J	510	561	1,580	72%
Beach 7	6	36.9	50.5	108.2	65.2	160	58%
Beach 8	9	92.1 J	204.2	71.9	123	302	58%
cPAH TEQ (µg/kg)							
Beach 1	3	362	111	35.3	169	600	101%
Beach 2	3	272	445	111	276	696	60%
Beach 3	5	197 J	83.7	20.7	100	325	90%
Beach 4	5	57.1	55.8	23.5	45.5	93.4	42%
Beach 5	9	357	41.9	3,050	1,150	5,310	144%
Beach 6	3	1,240	1,480	1,310	1,340	1,650	9%
Beach 7	6	38.3	38.5	52.4	43.1	63.4	20%
Beach 8	9	58.9	106	158	108	232	47%
Dioxin/Furan TEQ (ng/kg)							
Beach 1	3	1.39 J	1.96 J	1.47 J	1.61	2.38	19%
Beach 2	3	27.0 J	11.7 J	8.34 J	15.7	40.7	63%
Beach 3	5	4.62 J	8.19 J	0.306 J	4.37	14.3	90%
Beach 4	5	12.0 J	73.4 J	4.68 J	30.0	125	126%
Beach 5	9	4.40 J	6.41 J	5.07 J	5.29	7.87	19%

Beach Play Area	No. of Grab Samples per Composite	Composite Concentrations			Summary Statistics		
		Composite 1	Composite 2	Composite 3	Mean	95UCL[a]	CV
Beach 6	3	8.86 J	21.7 J	9.16 J	13.2	31.7	56%
Beach 7	6	1.87 J	2.24 J	2.27 J	2.13	2.69	10%
Beach 8	9	2.92 J	4.08 J	5.15 J	4.05	6.86	28%
Arsenic (mg/kg)							
Beach 1	3	4.93 J	16.0 J	23.2 J	14.7	37.9	63%
Beach 2	3	55.3 J	32.8 J	46.1 J	44.7	73.2	25%
Beach 3	5	4.60	2.96	4.48	4.01	6.31	23%
Beach 4	5	8.51 J	6.14 J	4.08 J	6.24	11.8	36%
Beach 5	9	5.52 J	12.4 J	8.31 J	8.74	17.5	40%
Beach 6	3	68.1	28.8	37.0	44.6	96.8	47%
Beach 7	6	4.95	4.78	6.60	5.44	7.97	18%
Beach 8	9	6.93	10.1	6.12	7.72	13	27%

[a] 95UCLs were calculated using Chebyshev's Inequality (n = 3 all areas). These estimates did not use the field replicates collected at Beaches 1 and 6.

95UCL – 95% upper confidence limit (on the mean)
cPAH – carcinogenic polycyclic aromatic hydrocarbon
CV – coefficient of variation
J – estimated concentration
N – tentative identification
PCB – polychlorinated biphenyl
SD – standard deviation
TEQ – toxic equivalent

B2.2.2.2 Evaluation of sampling variance

Variability measured in the beach play area sediment composites included spatial heterogeneity on two scales: a small, localized scale as heterogeneity from within the same sampling hole (measured with two field replicates), and a large scale as heterogeneity throughout each beach area sampled (measured with the three beach-wide composites). A VCA (Section B1.3) was used to quantify the small-scale spatial variance (i.e., the differences between replicate field samples) relative to the total variance within a beach play area (Table B2-7). This evaluation may be considered exploratory because of the limited number of replicates (i.e., two field replicates within each of the three beach-wide composites). However, this evaluation is useful for interpreting the current dataset, as well as providing information for modifying future sampling efforts.

Table B2-7. Results of VCA for intertidal sediment composite samples from beach play areas.

Variance Source	Degrees of Freedom	Sum of Squared Error	Mean Squared Error	Variance Component	% of Total for Variance Components
Total PCBs					
Beach 1 (CV = 109%)[a]	-	-	-	-	-
Total observed	2.1	na	na	23,225	100
Among composites	2	90,174	45,087	21,862	94
Within composite locations[b]	3	4,090	1,363	1,363	6
Beach 6 (CV = 67%)	-	-	-	-	-
Total observed	2.0	na	na	139,563	100
Among composites	2	555,072	277,536	137,973	99
Within composite locations[b]	3	4,769	1,590	1,590	1
cPAH TEQ					
Beach 1 (CV = 156%)	-	-	-	-	-
Total observed	3.5	na	na	301,767	100
Among composites	2	861,373	430,686	128,920	43
Within composite locations[b]	3	518,541	172,847	172,847	57
Beach 6 (CV = 136%)	-	-	-	-	-
Total observed	4.7	na	na	10,327,609	100
Among composites	2	21,589,026	10,794,513	466,904	5
Within composite locations[b]	3	29,582,115	9,860,705	9,860,705	95
Dioxins/furans TEQ					
Beach 1 (CV = 44%)	-	-	-	-	-
Total observed	4.8	na	na	0.8	100
Among composites	2	1.2	0.6	0[c]	0[c]
Within composite locations[b]	3	2.4	0.8	0.8	100

Variance Source	Degrees of Freedom	Sum of Squared Error	Mean Squared Error	Variance Component	% of Total for Variance Components
Beach 6 (CV = 57%)	-	-	-	-	-
Total observed	4.8	na	na	88	100
Among composites	2	87	43	0[c]	0[c]
Within composite locations[b]	3	264	88	88	100
Arsenic					
Beach 1 (CV = 52%)	-	-	-	-	-
Total observed	4.8	na	na	64	100
Among composites	2	78	39	0[c]	0[c]
Within composite locations[b]	3	191	64	64	100
Beach 6 (CV = 47%)	-	-	-	-	-
Total observed	2.8	na	na	360	100
Among composites	2	1,205	602	242	67
Within composite locations[b]	3	353	118	118	33

[a] CV is among all 6 composite samples.

[b] The variability is among three pairs of field replicates composed of sediment taken from the same holes as the primary samples.

[c] Negative variance component estimate is set to 0 (see Section B1.3).

cPAH – carcinogenic polycyclic aromatic hydrocarbon
CV – coefficient of variation
na – not applicable
PCB – polychlorinated biphenyl
TEQ – toxic equivalent
VCA – variance components analysis

Two field replicates were collected at Beaches 1 and 6. At each sampling location in these two areas, sediment from each hole was placed in two 16-oz jars (rather than one) for the field replicates. The field replicates were composited following the same methods and using the same locations as the original beach composite samples. Laboratory triplicates were analyzed for cPAHs only in composite sample 1 from Beach 1, so analytical variance was not included in the VCA.[12] Only the first cPAH TEQ reported for the primary composite sample 1 from Beach 1 was used in the following analysis.

For PCB Aroclors, the variability among field replicates was low (≤ 6% of total), indicating relative consistency among grab samples taken from the same holes.

For the other analytes, small-scale spatial variability was 95% or more of the total variance for the two beaches, with two exceptions: cPAHs at Beach 1 had small-scale spatial variability that contributed 57% to the total, and arsenic at Beach 6 had small-scale spatial variability that contributed 33% to the total.

The inference that can be made about the variance components in this dataset is limited due to the small dataset. The variance estimates for small-scale spatial variability in this

[12] The CVs for the laboratory replicates of composite 1 ranged from 5 to 12% for the individual cPAHs; these values were well within the laboratory analytical precision limit of 35%.

assessment are balanced but based on only two field replicates for each of the three composite samples. The conclusion that can be cautiously drawn from the results shown in Table B2-7 is that small-scale spatial variability contributes most of the total variance observed for some analytes. This was true for dioxins/furans at both beaches and for arsenic and cPAHs at least at Beach 1. However, for total PCBs, small-scale variability contributed ≤ 6% of the total variance. The total variance for cPAHs is high, with CVs ≥ 136% for the two beaches, and most of this variance is small-scale spatial variability (> 57%, Table B2-7).

The influence of small-scale spatial variability (variance between field replicates at each composite location) on the estimated results was evaluated by comparing the mean and 95UCL results both with and without the field replicates at Beaches 1 and 6 (Table B2-8). The calculations excluding the field replicates used only the three primary samples and Chebyshev's inequality for the 95UCL with two degrees of freedom. The calculations including the field replicates used all six results, first averaging the two replicates for each composite, and then calculating the 95UCL with two degrees of freedom using Chebyshev's inequality. Consistent with the VCA, when small-scale spatial variability was found to be a small percentage of the total (i.e., PCBs at both beaches and arsenic at Beach 6), the difference between the two 95UCLs was minimal. The most variable results for 95UCLs with and without field replicates were observed for cPAH TEQs. This variability was not widespread: Field replicates at Beach 1 had high variance between replicates from only one of the composites (LDW18-IT45-B1-Comp1), while at Beach 6 the variance was high between all three pairs of field replicates.

Table B2-8. Effect of field replicates on means and 95UCLs at Beaches 1 and 6

COC	Statistic	Beach 1		Beach 6	
		No Field Replicates[a] (n=3, Degrees of Freedom=2)	With Field Replicates (n=6, Degrees of Freedom=2)	No Field Replicates (n=3, Degrees of Freedom=2)	With Field Replicates (n=6, Degrees of Freedom=2)
Arsenic (mg/kg)	Mean	14.7	15.32	44.6	40.2
	SE	5.31	2.54	12.0	10.0
	95UCL[b]	37.9	26.4	96.8	83.9
Total PCB Aroclors (ug/kg)	Mean	120	140	561	554
	SE	74.5	86.7	234	215
	95 UCL	445	518	1582	1491
cPAH TEQ (ug/kg)	Mean	169	336	1343	2368
	SE	98.7	268	71.3	1341
	95UCL	600	1504	1654	8214
Dioxin/furan TEQ (ng/kg)	Mean	1.61	2.04	13.2	16.5
	SE	0.178	0.318	4.23	2.69
	95UCL	2.38	3.42	31.7	28.3

[a] When field replicates were excluded, only the three primary composite samples were used. When field replicates were included, the mean of the two composite samples with each sample ID (sample, and sample-FD) were averaged, and summary statistics were calculated from the three composite means per beach.

[b] The 95UCL for beaches were calculated using Chebyshev's Inequality (Equation 3, Section B2.2.2.1), with n = 3 in all areas.

95UCL – 95% upper confidence limit (on the mean)
COC – contaminant of concern
cPAH – carcinogenic polycyclic aromatic hydrocarbonID – identification
PCB – polychlorinated biphenyl
QAPP – quality assurance project plan
SE – standard error (of the mean)
TEQ – toxic equivalent

Pre-Design Studies results indicated that concentrations of total PCBs were significantly below the risk-based threshold concentration at all beaches, even with the conservative Chebyshev 95UCL. However, arsenic, cPAHs, and dioxins/furans exceeded the cleanup values at several beaches. Following any remediation, mean and variances are expected to be lower. The information regarding sources of variance obtained during the Pre-Design Studies, in addition to the spatial extent of any active remediation in each of the beach play areas, will be used to develop an appropriate long-term monitoring sampling plan for potential beach play area sediments.

B3 Surface Water

This section provides statistical details regarding the interpretation of the surface water data, as presented in Section 3 of the main report. Surface water grab samples were summarized in the main report, and no further discussion of these data is needed in this appendix. However, the C_{free} of total PCBs from the passive samplers are discussed herein with respect to mean, variance, and distribution.

B3.1 Distribution of Passive Sampler Results

During the development of the Work Plan (Windward and Integral 2017), the passive sampler study design was developed using the most recent passive sampler data from the LDW (i.e., passive sampler data from Apell and Gschwend 2017).

The Apell and Gschwend (2017) passive sampler data were limited to a single sample at three different locations. These data were insufficient to adequately evaluate the distributional form. Consequently, the *a priori* power calculations for the Work Plan were based on untested assumptions about the distributional form of the data and used both the normal and log-normal distributions. The total PCB baseline dataset was sufficient (n = 35)[13] to investigate the distributional form, so it was evaluated graphically using normal probability plots and formally using GOF tests (Section 1.1).

[13] Nine passive sampler replicates were deployed at each location in both baseline years (total n = 36). One passive sampler result was rejected from location PS1 in 2018 (Windward [in prep]), resulting in a total n = 35 for the baseline passive sampler dataset.

The normal probability plot for the station residuals[14] (Figure B3-1) indicated that data were approximately normally distributed; the Shapiro-Wilk GOF test did not reject normality (p=0.60). Consequently, a parametric ANOVA model may be used to assess these data.

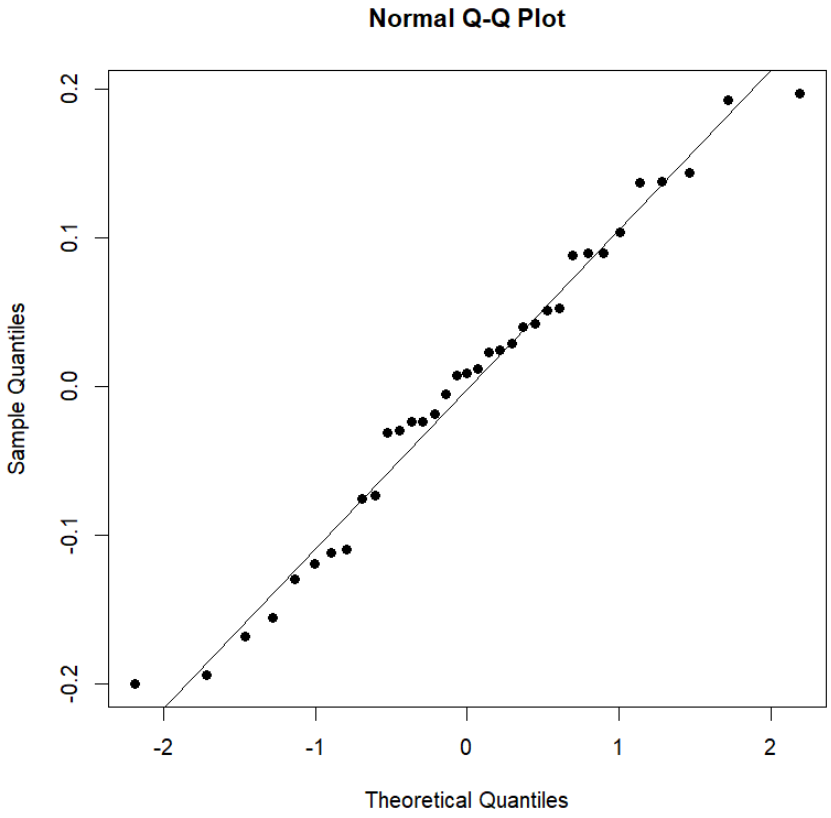


Figure B3-1. Normal probability plot of station year residuals for the baseline passive sampler dataset (n=35)

B3.2 EVALUATION OF SAMPLING VARIANCE

This passive sampler dataset provides a nearly balanced sampling design, with eight or nine replicates at both locations in both years. A balanced design, one with equal replication in each year for each station, makes the sums of squares in the ANOVA model additive (i.e., the individual sums of squares add up to the total), which provides the cleanest interpretation of the significance of the factors in the ANOVA model and the VCA. Balance was achieved by using the average of the eight replicates from

[14] Residuals are the individual observations minus the station year mean. The station year residuals have a common mean (zero), which allows results from the two stations and from each of the two years to be pooled to evaluate the shape and variance of these data, without the result being influenced by differences in the means from location or year.

location PS1 in 2018 as an estimate of the response for the rejected replicate at that station and year.

The statistical significance of Location and Year effects was tested using a two-factor crossed ANOVA model, including an interaction term (Location × Year) that allowed for the possibility that the two locations did not respond similarly over time (Table B3-1). The interaction term was not statistically significant ($p = 0.25$), indicating that the year-to-year differences observed at the two stations were similar (the temporal decrease at Station 2 [Sea Freeze] was 0.042 ng/L greater than at Station 1 [South Park Bridge] with a 95% confidence interval [0.039, 0.045]). The "location" row in Table B3-1 summarizes the differences between stations averaged between the two years; the average difference between stations (0.029 ng/L, 95% confidence interval [0.027, 0.032]) was not statistically significant ($p = 0.41$). The "year" row in Table B3-1 summarizes the differences between years averaged between the two stations; the average decrease from 2017 to 2018 (0.265 ng/L, 95% confidence interval [0.262, 0.268]) was highly statistically significant ($p < 0.001$).

Table B3-1. ANOVA table for comparison of total PCBs (ng/L) in passive samplers between two locations and two baseline years (2017 and 2018)

Source of Variance	Degrees of Freedom	Sum of Squares	Mean Squares	F Statistic[a]	p-Value
Year	1	0.6326	0.6326	55.553	<<0.001
Location	1	0.0078	0.0078	0.592	0.413
Location × year	1	0.0160	0.0160	1.317	0.245
Residuals	32	0.3644	0.0114	-	-

The F statistic is the ratio of appropriate mean squares, which is used to assess the significance of the source of variance; significance is indicated by the p-value.

ANOVA – analysis of variance

PCB – polychlorinated biphenyl

Using a VCA (Section 1.3), a relative comparison of the variance among location and year to the residual variability was made. The variance components for the total PCB passive sampler dataset are summarized in Table B3-2. The variability between locations was effectively 0% of the total, and the residual variability among replicate samplers was 25%. Most of the variability (i.e., 74% of the total) was between years.

Table B3-2. Results of VCA for total PCB passive sampler data

Variance Source	Degrees of Freedom	Sum of Squared Error	Mean Squared Error	Variance Component	% Total for Variance Components
Total observed	1.7	na	na	0.046	100%
Location	1	0.008	0.008	0[a]	0[a]
Year	1	0.633	0.633	0.034	74%
Location × year	1	0.016	0.016	0.001	1%
Residual	32	0.364	0.011	0.011	25%

Note: Imbalance in the design was corrected by using the average of the other eight replicates for the rejected replicate from station PS1 in 2018.

[a] Negative variance component set to zero (see Section 1.3).

na – not applicable

PCB – polychlorinated biphenyl

VCA – variance components analysis

From these results, the two different locations contribute very little to the variability in the dataset; the variability between locations accounts for essentially 0% of the total variance observed. The two passive sampler locations were selected to provide spatial coverage of the LDW (i.e., one location further downstream at RM 1.9 and one location further upstream at RM 3.3). The concentrations were very similar in both years, with differences between concentrations at the two locations averaging 0.029 ng/L (95% confidence interval = 0.0267, 0.0322). In contrast, the annual differences in concentrations constituted a large percentage of the variability of this dataset (74%, Table B3-2). The residual variability among the replicate samplers was relatively low (25%) and, most importantly, was very consistent between years and locations, suggesting that these samplers had high precision. From the evaluation of these data, temporal variability was much greater than spatial variability between the two passive sampler locations.

B3.3 POWER AND SAMPLE SIZE

The statistical approach for comparing the passive sampler data between baseline and future monitoring events is expected to be on a station-by-station basis. The variance estimate used in the *a priori* power analysis during Work Plan (Windward and Integral 2017) development was derived from the most recent passive sampler results from the LDW at that time (i.e., single replicate observations from each of three locations (Apell and Gschwend 2017)). Using residual variability from the recent baseline dataset, the power analysis was updated to assess the expected MDD between baseline and future monitoring events. The variance estimates from the recent baseline dataset are summarized in Table B3-3.

Table B3-3. Summary statistics for sum of PCB congeners from PE samplers deployed in the LDW

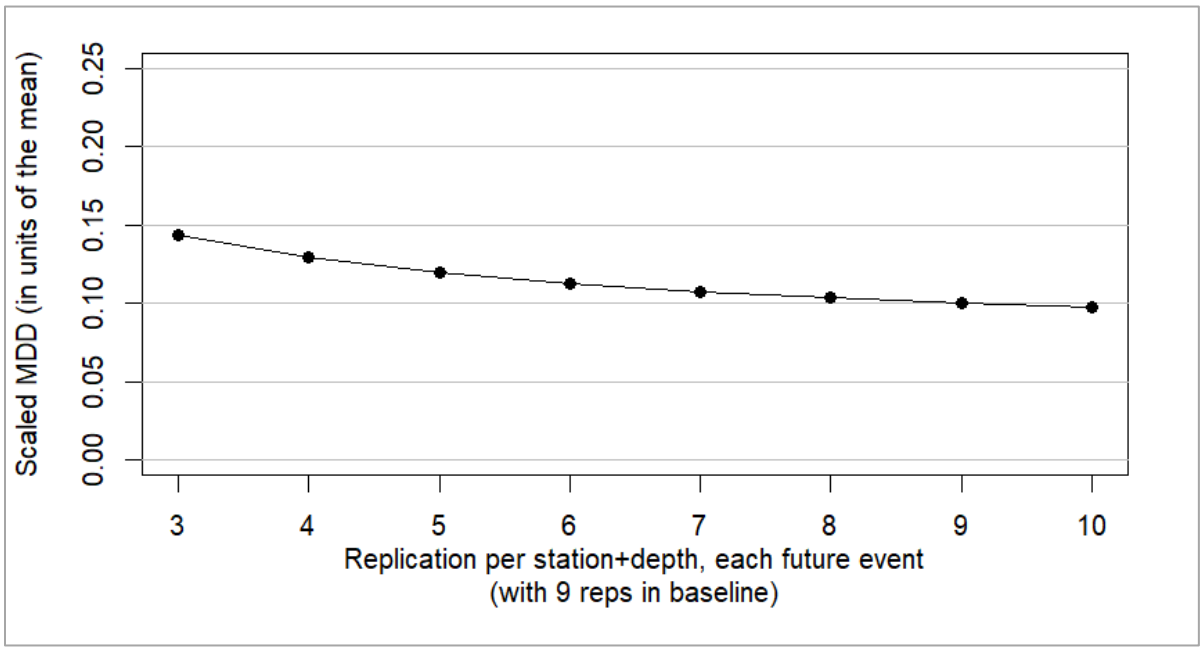
Summary Statistic	Pre-Design Studies Dataset	
	2017	2018
Sample size	18 (9 reps per station)	17 (8 at PS1 and 9 at PS2)
C_{free} total PCB mean concentration ($\bar{x}$) (ng/L)	1.26 (1.25 at PS1 and 1.26 at PS2)	0.99 (1.03 at PS1 and 0.96 at PS2)
SD for C_{free} total PCBs (ng/L)	0.115[a] (0.101 at PS1 and 0.128 at PS2)	0.101[a] (0.115 at PS1 and 0.086 at PS2)
CV = SD / $\bar{x}$	9.2%[b]	10.1%[b]

[a] The combined SD values reported for the Pre-Design Studies baseline samples are the residual SEs across both stations within each sampling year.

[b] The CVs reported for Pre-Design Studies baseline data use the values combined across the two stations.

CV – coefficient of variation
LDW – Lower Duwamish Waterway
PCB – polychlorinated biphenyl
PE – polyethylene
SD – standard deviation
SE – standard error (of the mean)

The SDs within the Pre-Design Studies passive samplers were very similar for the two baseline years (0.1 ng/L), and were low relative to the mean (CV of < 10%). Replicate variability was low for these passive samplers due to a combination of the extended exposure period (which avoided measuring the variance induced from short-term temporal fluctuations) and low instrument measurement error. Using a mean concentration of total PCBs for the Pre-Design Studies samples (2017–2018) of 1.125 ng/L, and a CV of approximately 10%, the MDD for a comparison between baseline and future monitoring is approximately 0.1125 ng/L for nine replicates (Type I and Type II errors = 0.10; comparison using a nested ANOVA model with two years nested within each study period [baseline and future] at a single location). Because of this very low residual variability, reducing the number of passive sampler replicates in future monitoring events to as few as three would still be expected to result in very tight MDDs of less than 15% of the mean (Figure B3-2). Five passive sampler results in future years would allow sufficient replicates to confirm the normality of the data and still achieve a low MDD (approximately 12%) for comparisons to baseline.



Note: Assumes a parametric 2-tailed *t*-interval testing for the difference of means between baseline (2 years) and future (2 years). Types I and II errors are both set at 10%. The CV value of 0.10 was observed during baseline sampling. Uses 9 replicates per station + depth during baseline sampling event. MDD (on the y-axis) is expressed as a percent of the baseline mean.

Figure B3-2. Relationship between replication within each station/depth for future sampling event and scaled MDD

B4 Fish and Crab Tissue

This section provides statistical details regarding the interpretation of the fish and crab tissue data, as presented in Section 4 of the main report.

B4.1 Influence of Non-Detects

Within the baseline fish and crab tissue datasets, data below detection had a noticeable influence only on the dioxin/furan results for the graceful crab edible meat samples. These non-detected dioxin/furan compounds introduced uncertainty to the calculated TEQ and affected whether the sample result was above or below the TTL in 6 of the 12 samples. Six graceful crab edible meat samples that had a dioxin/furan TEQ below the TTL using ½ the reporting limit (RL) for the non-detected compounds would have TEQs greater than the TTL if the full RL was used instead (with a TTL exceedance of 15% or less).

None of the other datasets were notably affected by non-detects. Total PCBs were calculated as the sum of detected Aroclors. Individual cPAH compounds were not detected in any of the crab tissue samples, and cPAHs were not analyzed in fish tissues because of the ability of fish to metabolize cPAH compounds. For dioxins/furans in tissues for which target tissue levels (TTLs) were available (i.e., graceful crab whole body and edible meat, and English sole whole body), the non-detected compounds notably affected only the graceful crab edible meat samples as noted above.

B4.2 95UCL Calculations

Fish and crab tissue DQO 1 required that a 95UCL for the site-wide mean be calculated for each tissue type. The sampling approach used a stratified design to account for possible differences of mean and variability in composite tissue concentrations across reaches and subreaches. As appropriate for the stratified sampling design, the site-wide mean was calculated as a stratified mean (i.e., a grand mean across strata with equal weights per stratum). For a UCL for a stratified mean, it is the distribution of the data within each stratum that is relevant; furthermore, if the data within individual strata are normal, then the mean of those strata means is also normal (because any linear combination of normal random variables is also normally distributed). Because of the relatively small sample sizes within each reach or subreach, residuals[15] within each reach or subreach were combined for greater power of the distributional test, and methods (described in Section 1.1) were used to identify the best distributional form for each COC and tissue type. The normal distribution was preferred for this stratified model; in all cases, the normal distribution provided a reasonable fit to the data, and no outliers were present (Table B4-1 and Figures B4-1 and B4-2). Total PCB Aroclors in several tissue datasets (i.e., English sole whole body [calculated], shiner surfperch, and

[15] Goodness-of-fit was applied to the residuals from a stratified model (i.e., the differences between each composite value and the mean for all samples from the same LDW river reach).

graceful crab whole body [calculated]) showed some deviations relative to the normal distribution. However, these datasets passed the GOF tests for normality, and the general symmetry and lack of extreme values within these datasets indicate acceptability of the normal distribution for calculating the 95UCLs.

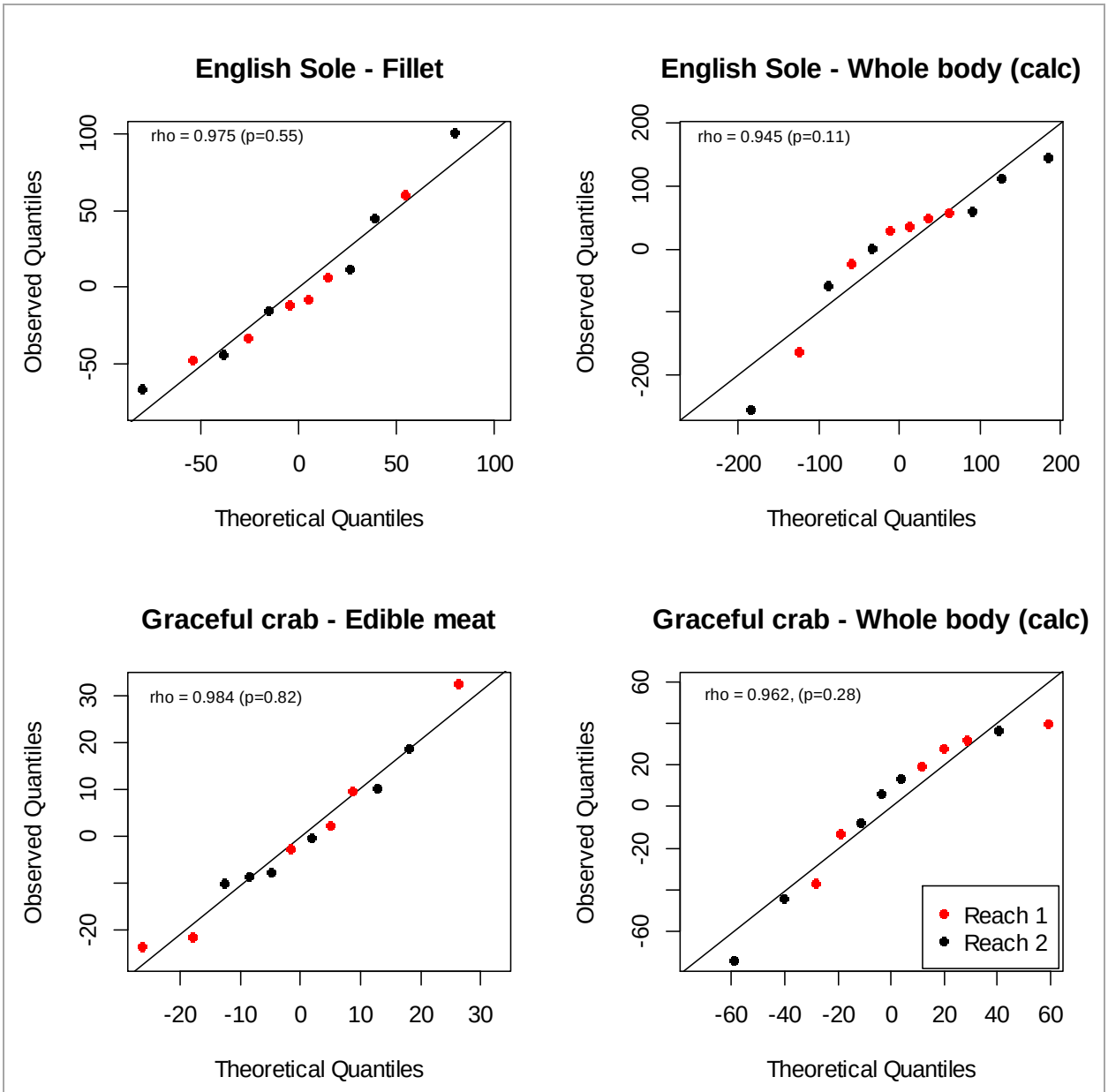
Table B4-1. GOF and CV summary for COCs in baseline fish and crab tissues

COC	Species and Tissue Type	Normal PPCC[a]	p-Value[b]	CV
Total PCBs (sum of Aroclors) (ug/kg ww)	English sole – fillet	0.975	0.55	0.20
	English sole – whole body (calculated)	0.945	0.11	0.16
	graceful crab – edible meat	0.984	0.82	0.15
	graceful crab – whole body	0.962	0.28	0.15
	shiner surfperch – whole body	0.956	0.20	0.08
Dioxin/furan TEQ (ng/kg ww)	English sole - fillet	0.988	0.91	0.20
	English sole – whole body (calculated)	0.978	0.64	0.11
	graceful crab – edible meat	0.964	0.32	0.19
	graceful crab – whole body	0.986	0.87	0.16
	shiner surfperch – whole body	0.984	0.83	0.28

[a] PPCC for the normal distribution.

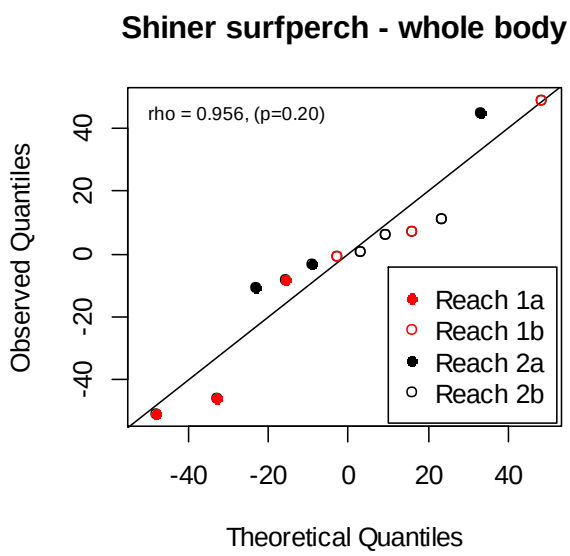
[b] p-value for the PPCC GOF test.

COC – contaminant of concern
CV – coefficient of variation
GOF – goodness-of-fit
PCB – polychlorinated biphenyl
PPCC – probability plot correlation coefficient
TEQ – toxic equivalent
TTL – target tissue level
ww – wet weight



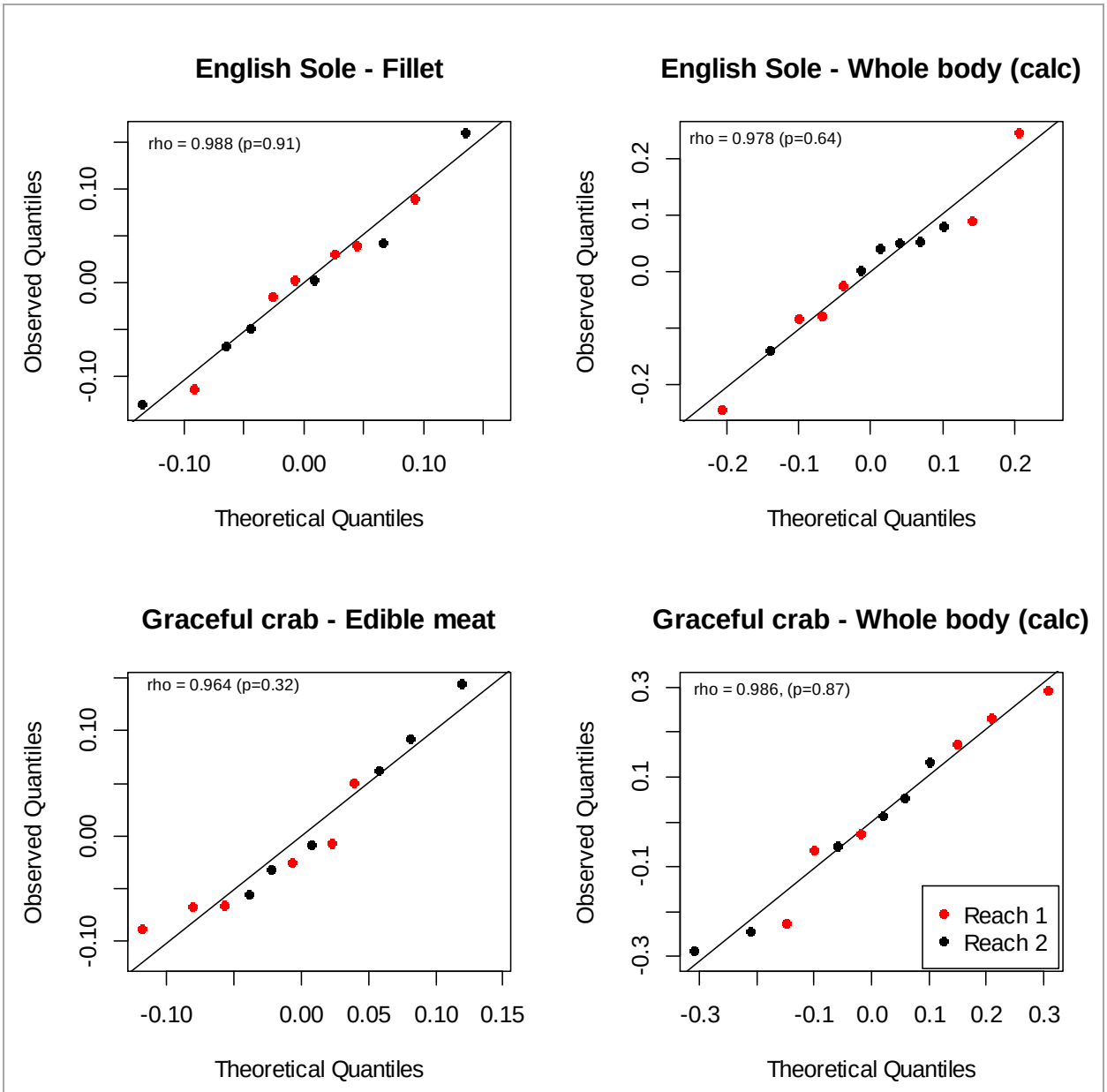
Results are shown for combined residuals from each LDW river reach, as appropriate for the stratified model. Residuals are color coded by reach.

Figure B4-1a. Normal probability plots of residuals by reach for baseline total PCB Aroclors (µg/kg ww) in English sole and graceful crab composite tissue samples



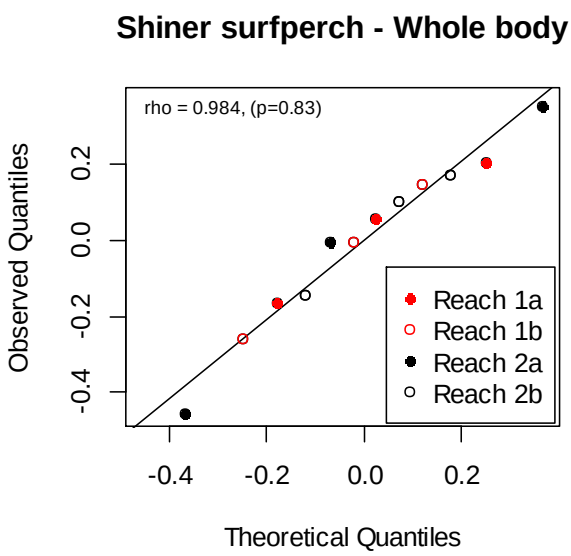
Results are shown for combined residuals from each LDW river reach, as appropriate for the stratified model. Residuals are color coded by reach.

Figure B4-1b. Normal probability plots of residuals by subreach for baseline total PCB Aroclors (µg/kg ww) in shiner surfperch composite tissue samples



Note: Results are shown for combined residuals from each LDW river reach, as appropriate for the stratified model. Data points are color coded by reach.

Figure B4-2a. Normal probability plots of residuals by reach for baseline dioxin/furan TEQ (ng/kg ww) in English sole and graceful crab composite tissue samples



Note: Results are shown for combined residuals from each LDW river reach, as appropriate for the stratified model. Data points are color coded by subreach.

Figure B4-2b. Normal probability plots of residuals by subreach for baseline dioxin/furan TEQ (ng/kg ww) in shiner surfperch composite tissue samples

B4.3 Statistical Comparisons Between 2017 Samples (Pre-Design Studies Dataset) and 2007 Samples

The temporal change in tissue concentrations between 2007 and 2017 was evaluated, to the extent allowed by the data, with statistical comparisons made between the average tissue concentrations of samples collected in 2007 as part of the LDWRI (Windward 2010) or the human health risk assessment (HHRA) (Windward 2007) and the average tissue concentrations of samples collected as part of the Pre-Design Studies in 2017. The designs from the two studies were generally similar, although there were differences in the number of fish/crab per composite, as well as the number of samples per location. These differences are noted in the following sections. In some cases, the data that were available precluded making site-wide comparisons, and statistical comparisons could only be made on a smaller spatial scale than the entire LDW.

B4.3.1 Total PCBs

Parametric ANOVA models were fit to each dataset, and standard residual diagnostic plots (i.e., normal probability plots, residuals vs. leverage values, and fitted values vs. standardized and unstandardized residuals) were used to assess the appropriateness of each model. If there were issues apparent in the diagnostic plots regarding the assumptions for the parametric ANOVA model, the statistical results were caveated.

The 95% confidence intervals on the differences of means between years were calculated with Welch's t-interval. Welch's interval accommodates differences in

variances and sample sizes for the two time periods. The sample sizes were small and unbalanced between the two time periods; in addition, slight differences in sampling locations and composite sizes between the surveys conducted in the two time periods mean that these results are approximations of the changes in mean concentrations over time.

In this section, multiple statistical comparisons are made, and with a Type I error (α) of 0.05, 5% of these comparisons may be statistically significant purely by chance—which is what the Type I error rate represents. The comparison-wise p-values were reported in the main text to indicate the strength of difference (or lack thereof) between the two study efforts. These results are meant to be informative of a pattern, not necessarily definitive statements regarding statistical significance.

B4.3.1.1 English sole fillet and whole-body tissues

Data for total PCB Aroclors in fillet and whole-body English sole samples for 2007 and 2017 are shown in Table B4-2. The sampling designs for the 2007 and 2017 samples were comparable—similar sampling areas were used and the sizes of the fish included in the composites were similar. However, the number of individual fish per composite was 5 in 2007 and 10 in 2017; as intended, this change resulted in a reduction in variance for the 2017 composite samples compared to 2007. The analysis used a crossed two-factor ANOVA; due to unequal sample sizes in each level of the design, a Type III ANOVA was used.[16] Differences in the mean concentrations over time and the associated 95% confidence intervals calculated using a Welch's t-interval are reported in Table B4-3. When the 95% confidence intervals contain zero, the differences between the two time periods are not statistically significant at the 95% confidence level. Wide intervals are an indication of large SEs due either to high sampling variability, small sample sizes, or both. The statistically significant effects are shaded in Table B4-3.

Table B4-2. Summary of total PCB results in English sole fillet and whole-body tissues for baseline and RI datasets

Dataset	English Sole Fillet		English Sole Whole Body			
	Average ± SD Total PCB Aroclor Concentration (µg/kg ww)	Sample Size	Average ± SD Total PCB Aroclor Concentration (µg/kg ww)	Sample Size	Average ± SD Total PCB Congener Concentration (µg/kg ww)	Sample Size
LDW RI 2007 (5 fish per composite)						
Reach 1	318 ± 113	6	609 ± 210	12	1,290 ± 407	4
Reach 2	403 ± 78	3	809[a]± 401	7	1,980 ± 1340	2
Mean of Reaches 1 and 2	361	9	709	19	1,640	6

[16] When a dataset does not have equal replication in every cell of the crossed-factor design, the mean squares for the significance tests may be calculated in different ways. A Type III ANOVA tests for the presence of a main effect, conditional on the other main effects and interactions. This approach is suitable when testing for main effects when interactions are present.

Dataset	English Sole Fillet		English Sole Whole Body			
	Average ± SD Total PCB Aroclor Concentration (µg/kg ww)	Sample Size	Average ± SD Total PCB Aroclor Concentration (µg/kg ww)	Sample Size	Average ± SD Total PCB Congener Concentration (µg/kg ww)	Sample Size
LDW Baseline 2017 (10 fish per composite)						
Reach 1	341 ± 65	6	888 ± 145	6	1,010 ± 174	3
Reach 2	177 ± 32	6	621 ± 85	6	606 ± 65	3
Mean of Reaches 1 and 2	259	12	754	12	808	6

[a] An extreme concentration was identified in this reach in 2007 (1,600 µg/kg ww). When excluded, the reach mean concentration was 677 ± 216 µg/kg ww, and the 2007 mean was 643 µg/kg ww. The influence of this sample was evaluated by comparing the two years without this sample (see Table B4-3).

LDW – Lower Duwamish Waterway
PCB – polychlorinated biphenyl
RI – remedial investigation
SD – standard deviation
ww – wet weight

Table B4-3. ANOVA table for comparison of total PCBs in English sole tissues between 2007 and 2017

Source	Degrees of Freedom	Sum of Squares	F Statistic	p-Value	Change Over Time[a] Difference of Means [95% Confidence Interval of Difference]	
					Reach 1	Reach 2
Total PCB Aroclors in English sole fillets						
(Intercept)	1	1,845,269	308.6	<<0.001		
Year	1	49,589	8.29	0.010	-23 [-145, 100]	226 [51, 400]
Reach	1	7,480	1.25	0.279		
Interaction	1	74,371	12.44	0.0026		
Residuals	17	101,655	-	-		
Total PCB Aroclors in English sole whole body						
(Intercept)	1	15,307,338	260.3	<<0.001		
Year	1	14,831	0.252	0.620	-279 [-460, -98]	188 [-183,560]
Reach	1	8,369	0.142	0.709		
Interaction	1	390,174	6.636	0.016		
Residuals	27	1,587,496	-	-		
Total PCB Aroclors in English sole whole body, excluding outlier from Reach 2 in 2007						
(Intercept)	1	13,388,849	406.32	<<0.001		
Year	1	85,250	2.59	0.120	-279 [-460, -98]	56 [-171, 284]
Reach	1	68,800	2.09	0.160		
Interaction	1	192,769	5.85	0.023		
Residuals	26	856,744	-	-		

Source	Degrees of Freedom	Sum of Squares	F Statistic	p-Value	Change Over Time[a] Difference of Means [95% Confidence Interval of Difference]	
					Reach 1	Reach 2
Total PCB congeners in English sole whole body[b]						
(Intercept)	1	16,877,289	57.13	<<0.001		
Year	1	1,938,729	6.56	0.034	283 [-332,898]	1374 [-10591,13340]
Reach	1	56,055	0.19	0.675		
Interaction	1	840,506	2.85	0.130		
Residuals	8	2,363,460	-	-		

Shading indicates a statistically significant main effect ($p \leq 0.05$) or potentially important interaction ($p \leq 0.25$). A higher alpha is used for interactions to avoid missing temporal patterns that may differ between reaches.

[a] Difference is between the means of the two datasets (i.e., 2007 mean minus 2017 mean). A positive value indicates a decrease in concentration over time. If the 95% confidence interval does not contain zero, the estimated change is significantly different from zero (at $\alpha = 0.05$). When an interaction between reach and year appears to be present ($p \leq 0.25$), temporal differences are summarized by reach.

[b] Normality and homogeneity of variances assumptions were challenged by the 2007 data from Reach 2. These results are approximate.

ANOVA – analysis of variance

PCB – polychlorinated biphenyl

B4.3.1.2 Shiner surfperch

Data for total PCB Aroclors in whole-body shiner surfperch samples for 2007 and 2017 are shown in Table B4-4. The sampling designs for the 2007 and 2017 samples were comparable—similar sampling areas were used, although the areas sampled during baseline were larger, and the sizes of the fish included in the composites were similar. However, the number of individuals per composite was 10 in 2007 and 15 in 2017; as intended, this change resulted in a reduction in variance for the 2017 composite samples compared to 2007. The analysis used a crossed two-factor ANOVA; due to unequal sample sizes in each level of the design, a Type III ANOVA was used. Differences in the mean concentrations over time and the associated 95% confidence intervals calculated using a Welch's t-interval are reported in Table B4-5. The statistically significant effects are shaded in Table B4-5.

Table B4-4. Summary of total PCB results in shiner surfperch whole-body tissues for 2007 and 2017 datasets

Dataset	Average ± SD Total PCB Aroclors Concentration (µg/kg ww)	Sample Size	Average ± SD Total PCB Congener Concentration (µg/kg ww)	Sample Size
LDW RI 2007 (10 fish per composite)				
Subreach 1a (T1)	268 ± 59	6	739 ± 332	2
Subreach 1b (T2)	415 ± 115	6	525 ± 174	4
Subreach 2a (T3)	763 ± 314	6	1,783 ± 961	2
Subreach 2b (T4)	315 ± 66	4	--	0

Table B4-4. Summary of total PCB results in shiner surfperch whole-body tissues for 2007 and 2017 datasets

Dataset	Average ± SD Total PCB Aroclors Concentration (µg/kg ww)	Sample Size	Average ± SD Total PCB Congener Concentration (µg/kg ww)	Sample Size
All subreaches combined	440	22	1,016	6
LDW Baseline 2017 (15 fish per composite)				
Subreach 1a	439 ± 48	3	496 ± 51	2
Subreach 1b	370 ± 48	3	405 ± 18	2
Subreach 2a	504 ± 11	3	551 ± 72	2
Subreach 2b	316 ± 7.5	3	333 ± 20	2
All subreaches combined	407	12	446	8

LDW – Lower Duwamish Waterway
PCB – polychlorinated biphenyl
SD – standard deviation
ww – wet weight

Table B4-5. ANOVA table for comparison of total PCBs in shiner surfperch samples

Source	Degrees of Freedom	Sum of Squares	F Statistic	p-Value	Change Over Time[a] Difference of Means [95% Confidence Interval of Difference] Reach 1	Reach 2
Total PCB Aroclors						
(Intercept)	1	5,517,293	239.64	<<0.001		
Year	1	8,491	0.37	0.549	-63 [-149, 24]	174 [-73, 421]
SubReach	3	478,912	6.93	0.001		
Interaction	3	187,968	2.72	0.065		
Residuals	26	598,611	-	-		
Total PCB Congeners[b]						
Year	1	847,689	5.53	0.051	182 [-203, 567]	1341 [-6955,9636]
Reach	3	1,125,345	2.45	0.149		
Interaction	2	741,746	2.42	0.159		
Residuals	7	1,072,710	-	-		

Shading indicates a statistically significant main effect ($p \leq 0.05$) or potentially important interaction ($p \leq 0.25$). A higher alpha is used for interactions to avoid missing temporal patterns that may differ between reaches.

[a] Difference is between the means of the two datasets (i.e., 2007 mean minus 2017 mean). A positive value indicates a decrease in concentration over time. If the 95% confidence interval does not contain zero, the estimated change is significantly different from zero (at $\alpha = 0.05$). When an interaction between reach and year appears to be present ($p \leq 0.25$), temporal differences are summarized by reach.

[b] Insufficient replication in all cells to estimate Type III ANOVA sums of squares; Type II model was used.

ANOVA – analysis of variance
PCB – polychlorinated biphenyl

B4.3.1.3 Graceful crab

Data for total PCB Aroclors in graceful crab tissue samples for 2007 and 2017 are shown in Table B4-6. Insufficient Dungeness crab samples were available for these two sampling years to conduct a statistical comparison, so only graceful crab was evaluated.[17] The sampling designs for the 2007 and 2017 samples were comparable—similar sampling areas were used, and the sizes of the crabs included in the composites were similar. However, the number of individuals per composite was 5 crabs in 2007 and 7 to 14[18] crabs in 2017. No graceful crab samples were available from Reach 2 in 2007, so the statistical comparisons between years were conducted using only samples from Reach 1. With equal replication in the two years, these data were analyzed using a single-factor standard ANOVA model to test for differences between the years (Table B4-7). The statistically significant effects are shaded in Table B4-7.

Table B4-6. Summary of total PCB Aroclor results in graceful crab tissues for the 2007 and 2017 datasets

Dataset	Edible Meat		Whole Body	
	Average ± SD Total PCB Concentration (µg/kg ww)	Sample Size	Average ± SD Total PCB Concentration (µg/kg ww)	Sample Size
LDW RI 2007 (5 crabs per composite)				
Reach 1 (T1 & T2)	41 ± 7.3	6	155 ± 54	6
LDW Baseline 2017 (7 crabs per composite) [see text]				
Reach 1	146 ± 15	6	319 ± 46	6
Reach 2	84 ± 18	6	192 ± 28	6
Reaches 1 and 2 combined	115	12	255	12

LDW – Lower Duwamish Waterway
na – not available
PCB – polychlorinated biphenyl
RI – remedial investigation
SD – standard deviation
ww – wet weight

Table B4-7. ANOVA table for comparison of total PCB Aroclor data in graceful crab samples (Reach 1 only)

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Statistic	p-Value	Change Over Time[a] Difference of Means [95% Confidence Interval of Difference]
Graceful crab edible meat (Reach 1 only)						
Year	1	33,444	33,444	231.9	3.03E-08	-106 [-122, -89]
Residuals	10	1,442	144	-	-	

[17] Graceful crab is more commonly available in the LDW than Dungeness crab, so graceful crab was collected for the purpose of trend evaluations.
[18] Seven crab were included in the graceful crab edible meat composites, and 14 graceful crab were included in the hepatopancreas composites.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Statistic	p-Value	Change Over Time[a] Difference of Means [95% Confidence Interval of Difference]
Graceful crab whole body (Reach 1 only)						
Year	1	80,688	80,688	31.79	0.000216	-164 [-229, -99]
Residuals	10	25,384	2,538	-	-	

Shading indicates a statistically significant effect ($p \leq 0.05$).

[a] Difference is between the means of the two datasets (i.e., HHRA mean minus baseline mean). A positive value indicates a decrease in concentration over time. If the 95% confidence interval does not contain zero, the estimated change is significantly different from zero (at $\alpha = 0.05$).

ANOVA – analysis of variance

HHRA – human health risk assessment

PCB – polychlorinated biphenyl

B4.3.2 Inorganic arsenic

The concentrations for inorganic arsenic in tissues reported in the HHRA were presented in Table 4-11 (main report) and graphically compared to the tissue from the baseline studies in Figure B4-3. Where the two datasets were suitably comparable, the HHRA and baseline datasets were statistically compared to estimate the temporal change in mean concentrations. The analysis used a crossed two-factor ANOVA; due to unequal sample sizes in each level of the design, a Type III ANOVA was used. Differences in the mean concentrations over time and the associated 95% confidence intervals calculated using a Welch's t-interval are reported in Table B4-8, where statistically significant effects are shown with shading.

Table B4-8. Comparison of mean inorganic arsenic concentrations in fish and crab tissues between HHRA and baseline datasets

Source	Degrees of Freedom	Sum of Squares	F Statistic	p-Value	Change Over Time[a] Difference of Means [95% Confidence Interval of Difference]	
					Reach 1	Reach 2
English sole whole body						
(Intercept)	1	0.152796	106.94	<<0.001	-	-
Year	1	0.020856	14.60	0.0015	-0.004 [-0.05, 0.04]	-0.13 [-0.19, -0.06]
Reach	1	0.019712	13.80	0.0019		
Interaction	1	0.018204	12.74	0.0026	-	-
Residuals	16	0.02286	-	-	-	-
Shiner surfperch whole body						
(Intercept)	1	0.065054	59.65	<<0.001	-	-
Year	1	0.00267	2.45	0.137	0.024 [-0.08, 0.12]	
Reach	1	0.000028	0.03	0.875		
Interaction	1	0.001484	1.36	0.261	-	-
Residuals	16	0.017451	-	-	-	-

Source	Degrees of Freedom	Sum of Squares	F Statistic	p-Value	Change Over Time[a] Difference of Means [95% Confidence Interval of Difference]	
					Reach 1	Reach 2
Graceful crab edible meat (reaches combined; insufficient data in 2007 for Reach 2)						
Year	1	0.0135	5.52	0.034	-0.067 [-0.102,-0.032]	
Residuals	14	0.0343	-	-	-	-
Graceful crab whole body (reaches combined; insufficient data in 2007 for Reach 2)						
Year	1	0.00075	0.422	0.526	-0.016 [-0.067, 0.036]	
Residuals	14	0.02495	-	-	-	-

Shading indicates a statistically significant main effect ($p \leq 0.05$) or potentially important interaction ($p \leq 0.25$). A higher alpha is used for interactions to avoid missing temporal patterns that may differ between reaches.

[a] Difference is between the means of the two datasets (i.e., 2007 mean minus 2017 mean). A positive value indicates a decrease in concentration over time. If the 95% confidence interval does not contain zero, the estimated change is significantly different from zero (at $\alpha = 0.05$). When an interaction between reach and year appears to be present ($p \leq 0.25$), temporal differences are summarized by reach.

HHRA – human health risk assessment

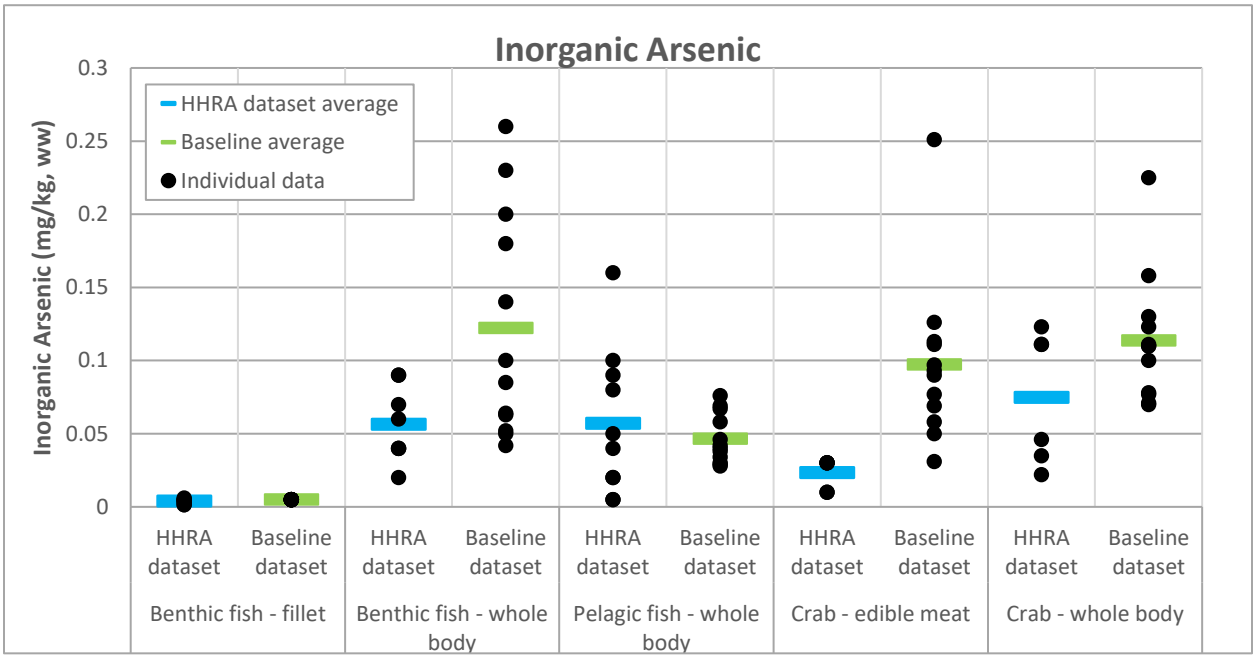


Figure B4-3. Comparison of inorganic arsenic concentrations in tissues in HHRA and baseline datasets

B4.4 Relationship Between PCB Aroclors and Congeners in Fish and Crab Tissues

A subset of the fish and crab tissue samples were evaluated for both PCB Aroclors and congeners (Table B4-8). Within each species and tissue type, the correlation between the two PCB estimates was evaluated for consistency of results. The slopes of the linear

regressions between the Aroclor and congener sums were significantly different from zero (adjusted p-value < 0.05, Table B4-8) for all tissues except English sole fillet. In addition, the regression slopes were all less than one, indicating that Aroclors under-predicted congeners. The estimated slopes for the graceful crab tissues were very close to one, indicating very good consistency between the two total PCB estimates in these tissues. The data with the ordinary least squares (OLS) regression lines are shown in Figure B4-4.

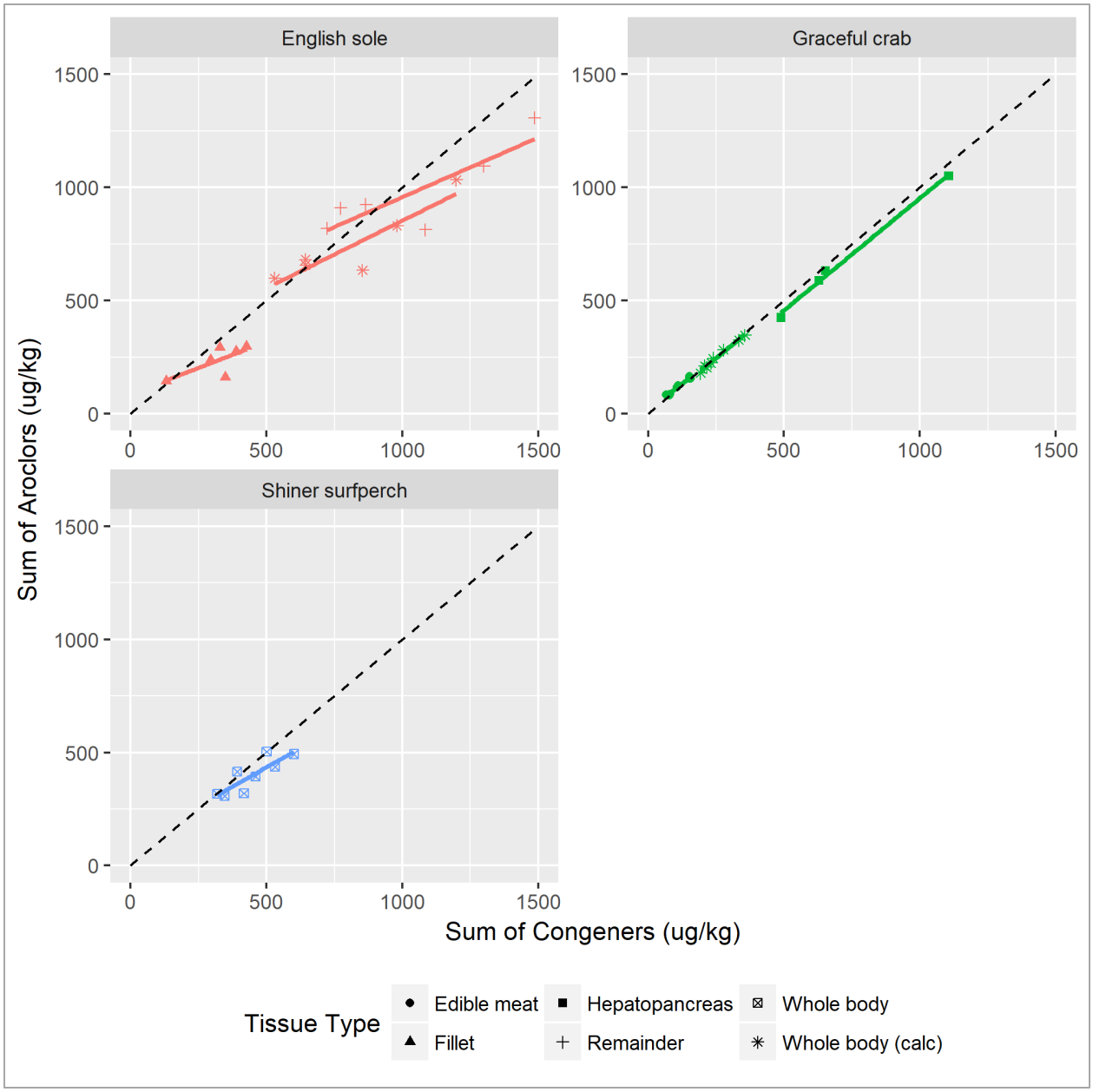
Table B4-8. Regression results between PCB Aroclors and congeners in baseline tissue samples of fish and crab

Species - Tissue Type	No. of Samples with Both PCB Aroclors and Congeners	Regression R^2	Slope of OLS Regression Line [95% Confidence Interval]	Adjusted p-Value for Linear Slope[a]
English sole - fillet	6	0.51	0.46 [-0.16, 1.09]	0.11
English sole - remainder	6	0.72	0.53[b] [0.07, 0.99]	0.04
English sole - whole body (calculated)	6	0.82	0.60[b] [0.21, 0.98]	0.02
Graceful crab - edible meat	8	0.97	0.93[b] [0.78, 1.09]	0.00
Graceful crab - hepatopancreas	4	0.99	0.99[b] [0.77, 1.21]	0.01
Graceful crab - whole body (calculated)	8	0.98	0.99[b] [0.86, 1.12]	0.00
Shiner surfperch - whole body	8	0.71	0.68[b] [0.25, 1.12]	0.01

[a] The p-values were adjusted to control the false discovery rate, or Type I error rate, among rejected hypotheses for the set of regressions on 8 tissue types. The adjusted p-values were calculated using the *p.adjust(method="BH")* function in R (R Core Team 2018).

[b] Indicates regression slope is significantly different from 0 (i.e., adjusted p-value < 0.05).

OLS – ordinary least squares



Note: Each species is shown in its own panel, and different symbols indicate the tissue type. The OLS linear regression lines are shown for each species-tissue type combination (see Table B4-8 for regression results). The black dashed line on each panel is the 1:1 line.

Figure B4-4. Plot of Aroclors vs. congeners for baseline fish and crab tissues

B4.5 POWER AND SAMPLE SIZE

Using the CV results from the Pre-Design Studies, the expected MDDs for comparison between baseline and future monitoring were calculated for total PCB Aroclors and dioxin/furan TEQ for each species and tissue type with TTLs (Table B4-9). For all tissue types, the baseline CVs were lower than the estimates used during Work Plan development. A consequence of a lower CV is increased statistical power for comparisons between baseline and future monitoring. The estimated MDDs ranged

from 10 to 25% of the baseline mean, indicating that the baseline design is statistically sufficient to detect meaningful changes in tissue concentrations.

Table B4-9. Expected MDDs for comparisons between baseline and future site-wide means of COCs in species/tissue types with TTLs

Chemical	Species and Tissue Type	Work Plan CV	Baseline Site-wide Mean	Baseline CV	Power Calculations (α = 0.10, power = 0.90)		
					MDD[a] as Conc.	MDD as % of Baseline Mean	Future Means Expected to be Significantly Less than Baseline Mean
Total PCBs (Aroclors) (µg/kg ww)	English sole – fillet	0.40	259	0.20	65	25%	< 194 µg/kg ww
	shiner surfperch – whole body	0.40	407	0.08	41	10%	<366 µg/kg ww
	graceful crab – edible meat	0.25	115	0.15	21	19%	< 94 µg/kg ww
	graceful crab – whole body	0.25	255	0.15	48	19%	< 207 µg/kg ww
Dioxin/furan TEQ (ng/kg ww)	English sole – whole body	0.40	1.18	0.11	0.16	14%	< 1.02 ng/kg ww
	graceful crab – edible meat	0.25	0.406	0.19	0.10	24%	< 0.306 ng/kg ww
	graceful crab – whole body	0.25	1.21	0.16	0.24	20%	< 0.97 ng/kg ww

[a] The MDD is the minimum detectable difference for a comparison between baseline and a future monitoring event, both using the baseline study design. The MDD calculations used a crossed ANOVA model, with sampling reach (or subreach, for shiner surfperch) crossed with year and total n = 12 in each year.

ANOVA – analysis of variance
COC – contaminant of concern
CV – coefficient of variation
MDD – minimum detectable difference
PCB – polychlorinated biphenyl
TTL – target tissue level
ww – wet weight

Dioxin/furan TEQ values in graceful crab edible meat and whole body (calculated) had 95UCLs that were less than the TTL (Table 4-4 in main report). Dioxin/furan TEQ datasets for both tissue types were found to be normally distributed, so the statistical power for a one-tailed, one-sample t-test (α = 0.05) comparing these data to the TTL was calculated. The power of the comparison to the TTL was > 99% for dioxin/furan TEQ in these tissue types.

B5 Clam Tissue

This section provides statistical details regarding the interpretation of the clam tissue data presented in Section 5 of the main report.

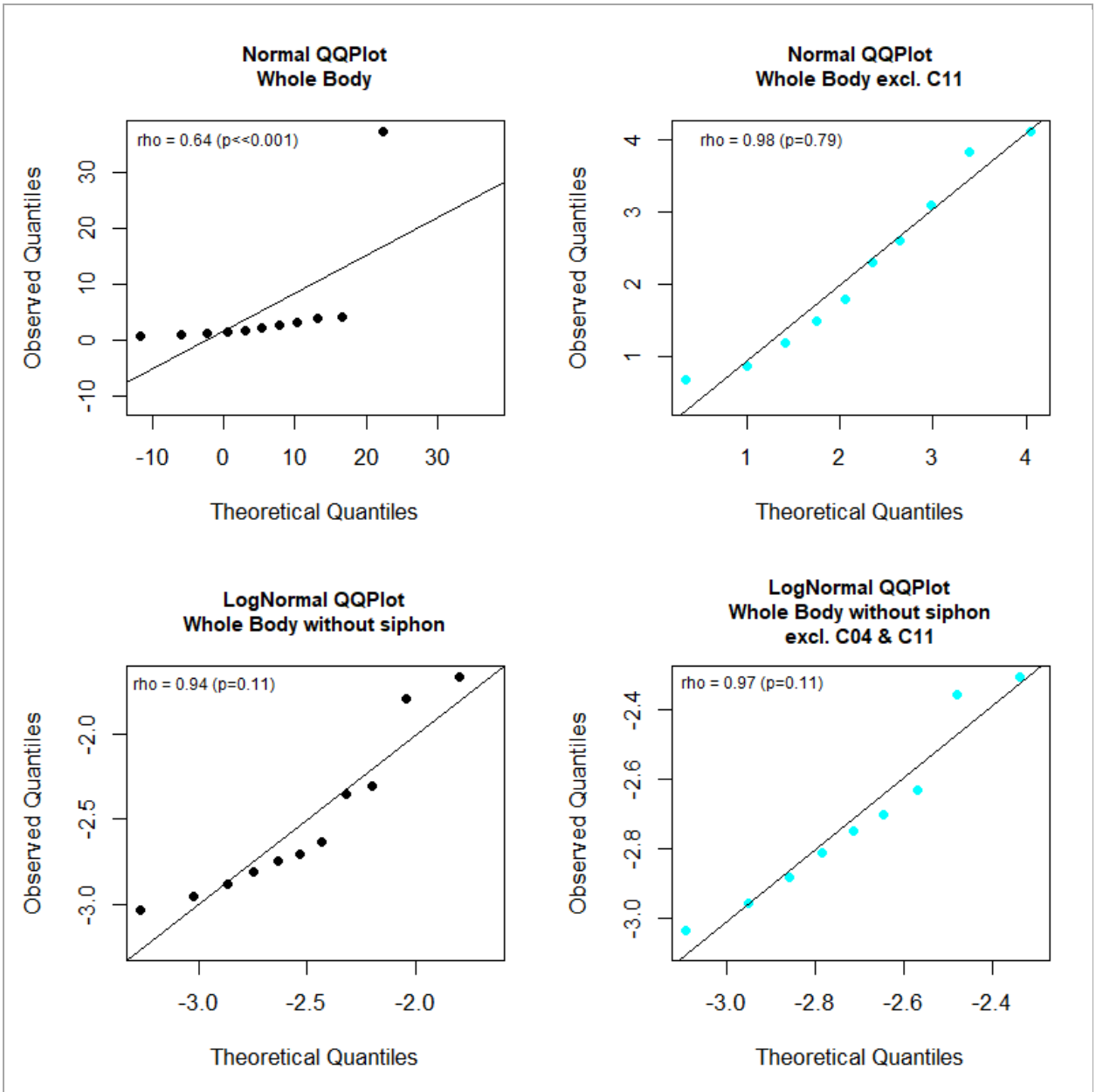
B5.1 Influence of Non-Detects

Total PCBs are calculated as the sum of detected Aroclors. For dioxins/furans, the contribution to the total from non-detected compounds ranged from 1 to 60% (using 0.5 method detection limit [MDL]). Individual cPAH compounds below detection contributed 0 to 81% of the total TEQ (using 0.5 MDL). The influence of non-detects on

the overall cPAH TEQ led to an analysis of several different treatments of the detection limits in calculation of the TEQ.

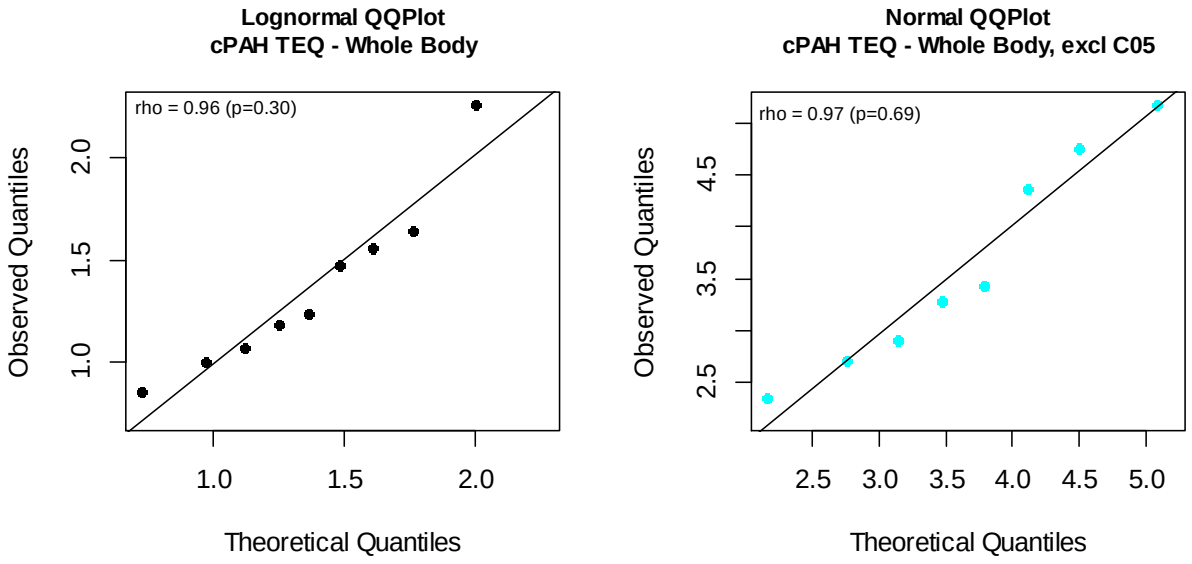
B5.2 95UCL CALCULATIONS

The clam tissue DQO 1 required that the 95UCL for the site-wide mean be established from this dataset for the four risk drivers. Following the methods described in Section 1.1, the best distributional form for each COC was identified in order to use the most appropriate result generated using ProUCL5.1 as the basis for the 95UCL. The best-fitting probability plots are shown in Figures B5-1 through B5-3, and results are summarized in Table B5-1.



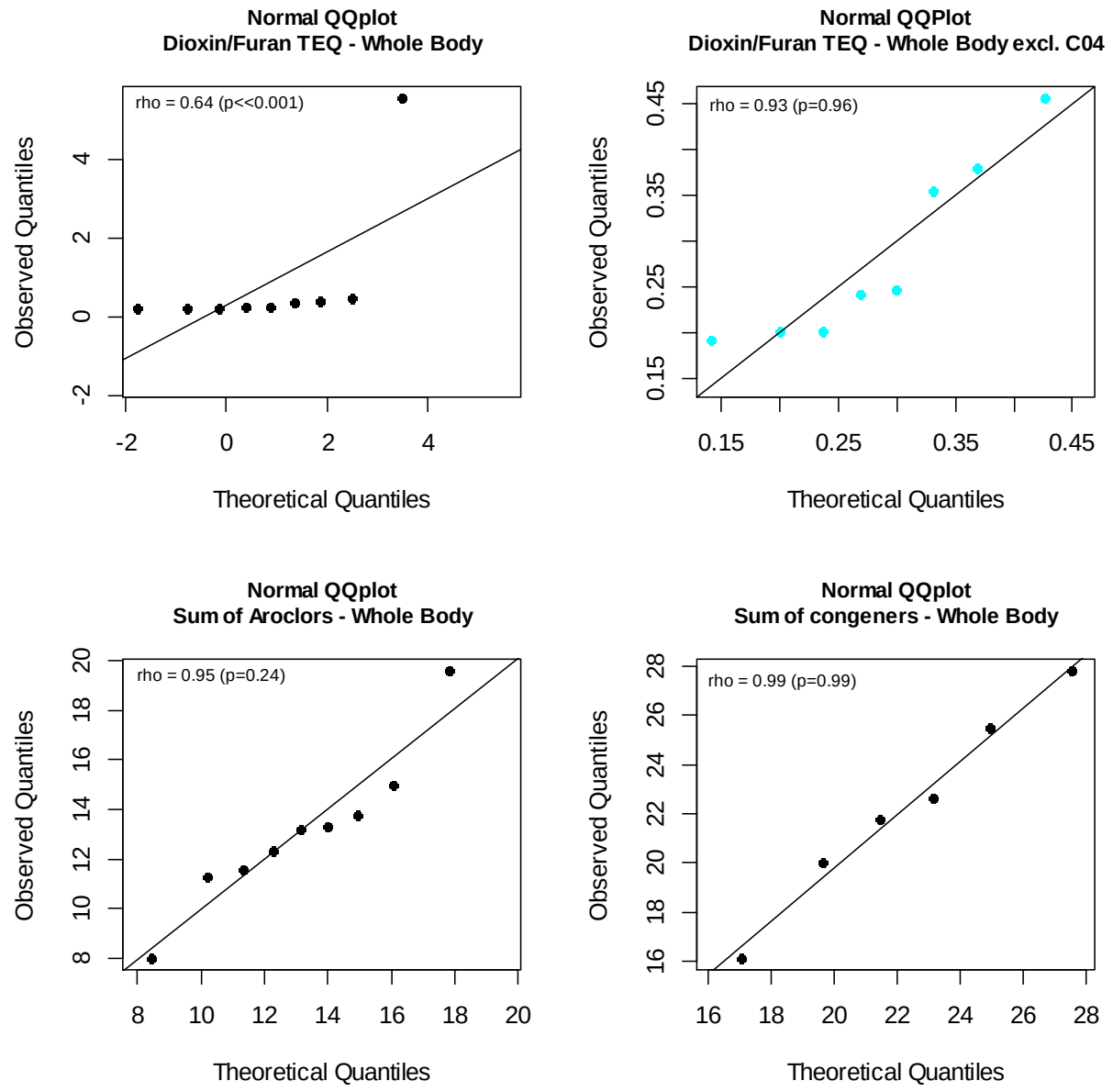
Note: Inorganic arsenic in whole-body (calculated) tissues for all data (top left) and excluding the outlier in area C11 (RM 3.8E) (top right). Inorganic arsenic in whole-body tissues without siphon skin for all data (bottom left) and excluding influential values from areas C04 and C11 (RM 1.5W and RM 3.8E) (bottom left). Plots with teal dots show the best-fitting distribution excluding the extreme and influential values that skewed the complete distribution.

Figure B5-1. Probability plots of inorganic arsenic (mg/kg ww) results in clam tissue composite samples



Note: cPAH TEQs in whole-body tissues for all data using full MDL for non-detects (top left), ½ MDL (top right), and zero MDL (bottom left). Results using full MDL and excluding the extreme value in area C05 (Slip 2) (bottom right). Plots with teal dots show the best-fitting distribution excluding the extreme and influential values that skewed the complete distribution.

Figure B5-2. Probability plot of cPAH TEQ (µg/kg ww) results in clam tissue composite samples (using the ultra-trace results)



Note: Dioxin/furan TEQs for all data (top left) and excluding the outlier in area C04 (Glacier Bay) (top right). Total PCB sum of Aroclors (bottom left) and sum of congeners (bottom right). Plots with teal dots show the best-fitting distribution excluding the extreme and influential values that skewed the complete distribution.

Figure B5-3. Probability plots for dioxin/furan TEQs (ng/kg ww) and total PCBs (µg/kg ww) in clam tissue composite samples

Table B5-1. Goodness-of-fit and variance statistics for risk drivers in clam tissue composite samples

Risk Driver	n	Best Fit Distribution	PPCC	p-Value	CV
Total PCBs (µg/kg ww)					
Total PCB Aroclors	9	normal	0.95	0.24	0.24
Total PCB congeners	6	normal	0.99	0.99	0.18
Dioxin/furan TEQ (ng/kg ww)					
All data	9	none	0.64	< 0.001	2.0
Excluding highest value from area C04 (Glacier Triangle)	8	normal	0.93	0.96	0.35
cPAH TEQ (µg/kg ww) [a]					
All data	9	lognormal	0.96	0.30	0.51
Excluding highest value from area C05 [Slip 2]	8	normal	0.97	0.69	0.28
Inorganic arsenic (mg/kg ww)					
Whole body (calculated) (all data)	11	none	0.64	< 0.001	1.98
Whole body (calculated) (excluding highest value from area C11 at RM 3.8E)	10	normal	0.98	0.79	0.54
Whole body without siphon skin	11	lognormal	0.94	0.11	0.54
Whole body without siphon skin – excluding highest values from areas C04 and C11	9	lognormal	0.97	0.11	0.46

[a] cPAH TEQs were calculated with the results of a re-analysis of the clam tissue samples using the ultra-trace modified method (EPA method 8270/1625).

cPAH – carcinogenic polycyclic aromatic hydrocarbon
CV – coefficient of variation
EPA – US Environmental Protection Agency
MDL – method detection limit
PCB – polychlorinated biphenyl
RM – river mile
TEQ – toxic equivalent
ww – wet weight

When individual elevated sample(s) were responsible for the skewness in a distribution, that dataset was also evaluated without the elevated sample(s). The elevated samples were collected from areas expected to be remediated, so the calculations that excluded outliers explored how the data may be expected to behave post-remediation, whereas the baseline 95UCLs are represented by the complete datasets.

B5.3 POWER AND SAMPLE SIZE

In the future, the sample size and variability for the clam tissue dataset will be determined by the number of clamming areas in which clams are present. The statistical power of the clam tissue dataset will be addressed when clam populations have recovered from active remediation, and estimates of variance can be assessed using the most recent data at that time.

B6 References

Apell JN, Gschwend PM. 2017. The atmosphere as a source/sink of polychlorinated biphenyls to/from the Lower Duwamish Waterway Superfund site. Environ Pollut 227:263-270.

DMMP. 2009. *OSV Bold* summer 2008 survey. Data report. The Dredged Material Management Program (DMMP) agencies: US Army Corps of Engineers, Seattle District, Seattle, WA; US Environmental Protection Agency, Region 10, Seattle, WA; Washington State Department of Natural Resources; and Washington State Department of Ecology, Olympia, WA.

Ecology. 2015. Sediment cleanup users manual II. Guidance for implementing the Cleanup Provisions of the Sediment Management Standards, Chapter 173-204 WAC. March 2015. Pub. no. 12-09-057. Toxics Cleanup Program, Washington State Department of Ecology, Olympia, WA.

EPA. 2014. Record of Decision. Lower Duwamish Waterway Superfund Site. US Environmental Protection Agency.

EPA. 2016. Statistical software ProUCL 5.1.00 for environmental applications for data sets with and without nondetect observations [online]. US Environmental Protection Agency, Washington, DC. Updated June 20, 2016. Available from: https://www.epa.gov/land-research/proucl-software.

Millard SP. 2013. EnvStats. An R Package for Environmental Statistics. Springer.

R Core Team. 2018. R: A language and environment for statistical computing [online]. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/.

Schuetzenmeister A, Dufey F. 2018. VPF: variance component program [online]. Updated July 18, 2018. Available from: https://cran.r-project.org/web/packages/VCA/index.html.

Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, NY.

Windward. 2007. Lower Duwamish Waterway remedial investigation. Baseline human health risk assessment. Prepared for Lower Duwamish Waterway Group. Windward Environmental LLC, Seattle, WA.

Windward. 2010. Lower Duwamish Waterway remedial investigation. Remedial investigation report. Final. Prepared for Lower Duwamish Waterway Group. Windward Environmental LLC, Seattle, WA.

Windward, Integral. 2017. Pre-design studies work plan. Lower Duwamish Waterway Superfund site. Final. Prepared for the Lower Duwamish Waterway Group for submittal to EPA Region 10 on August 28, 2017. Windward Environmental LLC and Integral Consulting Inc., Seattle, WA.

Windward. 2018a. Lower Duwamish Waterway baseline surface sediment collection and chemical analyses - quality assurance project plan. Final. Windward Environmental LLC, Seattle, WA.

Windward. 2018b. Lower Duwamish Waterway surface sediment data report. Draft final. Submitted to EPA October 9, 2018. Windward Environmental LLC, Seattle, WA.

Windward. [in prep]. Baseline surface water collection and chemical analyses data report. Draft. Submitted to LDWG November 13, 2018. Windward Environmental LLC, Seattle, WA.